Additional screening with ultrasound after negative mammography screening in women with dense breasts: a systematic review

External review against programme appraisal criteria for the UK National Screening Committee (UK NSC)

Draft report

Review Group: **Jacoby Patterson**

Chris Stinton

Lena Alkhudairy

Amy Grove

Pam Royle

Hannah Fraser

Hema Mistry

Payagalage Senaratne

Aileen Clarke

Sian Taylor-Phillips

Correspondence to: **Chris Stinton**

Populations, Evidence and Technologies

Division of Health Sciences

Warwick Medical School

University of Warwick

Coventry CV4 7AL

Tel: **02476 574 701**

Email: **C.Stinton@warwick.ac.uk**

Date completed: 02/03/2018

Funding Acknowledgement

This research was commissioned by the UK National Screening Committee. Sian Taylor-Phillips, Chris Stinton, Hannah Fraser, Lena Alkhudairy, Amy Grove and Aileen Clarke are supported by the National Institute for Health Research (NIHR) Collaboration for Leadership in Applied Health Research and Care West Midlands (NIHR CLAHRC WM). The views expressed in this publication are those of the authors and not necessarily those of the NHS, the National Institute for Health Research, the UK National Screening Committee, Public Health England or the Department of Health. Any errors are the responsibility of the authors. The authors have no conflicts of interest.

Expert Acknowledgement

We would like to thank Dr Sue Astley, Dr Eleanor Cornford, Dr Stephen Duffy, Dr Jacqui Jenkins, Dr Olive Kearins, Dr Paul Pharoah, Dr Nisha Sharma, Dr Matthew Wallis and Dr Louise Wilkinson for providing advice and input into this research.

Patient and Public Involvement Acknowledgement

We would like to thank Jane Whitehurst for her advice and input into this work especially the lay summary.

Special thanks go to Peter Chilton of the CLAHRC WM for his work on the diagrams.

Contents

Appendix 7 Criteria for appraising the viability, effectiveness and appropriateness of a screening programme 241

ABBREVIATIONS

ABUS	Automated whole breast ultrasound
AMSTAR	A measurement tool to assess systematic reviews
AUC	Area under the receiver operating characteristic curve
BI-RADS	Breast Imaging Reporting and Data System
BMI	Body mass index
BRCA	Breast Cancer gene
BSP	Breast Screening Programme
CC	Cranio-caudal
CCC	Concordance Correlation Coefficient
CHEERS	Consolidated Health Economic Evaluation Reporting Standards
DBT	Digital breast tomosynthesis
DCIS	Ductal carcinoma in situ
DM	Digital mammogram
DOR	Diagnostic Odds Ratio
ER/PR	Estrogen receptor/progesterone receptor
ES	Effect size
FFDM	Full-field digital mammography
FN	False negative
FP	False positive
GRRAS	Guidelines for Reporting Reliability and Agreement Studies
HER2	Human epidermal growth factor receptor type 2
HHUS	Hand-held ultrasound
HR	Hormone receptor
HRT	Hormone replacement therapy
ICC	Intraclass correlation coefficient
ICER	Incremental Cost-Effectiveness Ratio
IQR	Inter-quartile range
κ_w	Weighted kappa
LIBRA	Laboratory for Individualized Breast Radiodensity Assessment
LR+/-	Positive/negative likelihood ratio
MBTST	Malmo Breast Tomosynthesis Screening Trial
MLO	Medio-lateral oblique
MRI	Magnetic resonance imaging
NHSBSP	UK National Health Service Breast Screening Programme
NPV	Negative predictive value
NSC	National Screening Committee
OR	Odds ratio
PD	Percent density
PHE	Public Health England
PPV	Positive predictive value
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses

PROSPERO	The International Prospective Register of Systematic Reviews
QALY	Quality-adjusted life year
QAREL	Quality Appraisal of Diagnostic Reliability
QUADAS	Quality assessment tool for diagnostic accuracy studies
QUIPS	Quality in Prognostic Studies
RANZCR	The Royal Australian and New Zealand College of Radiologists
RCT	Randomised controlled trial
REA	Rapid evidence assessment
ROC	Receiver operating characteristic
RR	Relative risk
SD	Standard deviation
sDM	Synthetic digital mammogram
SR	Systematic review
SXA	Single energy x-ray absorptiometry
TN	True negative
TP	True positive
UK	United Kingdom
US	Ultrasound
USA	United States of America
USPTF	United States Preventive Task Force
VDG	Volumetric density grade

Executive summary

Background: The NHS Breast Screening programme screens women aged 50-70 using mammography every 3 years, with no routine measurement or reporting of mammographic breast density. Some other countries report mammographic breast density to women attending screening, as the dense breast parenchyma may obscure cancer on a mammogram and density may itself be a risk factor for developing cancer. Others offer additional ultrasound testing for women with mammographically dense breasts.

Objectives: To determine the balance of benefits and costs of measuring breast density on mammography, and offering women with dense breasts supplemental ultrasound screening. The United Kingdom (UK) National Screening Committee (NSC) criteria for appraising screening programmes state that there should be a validated screening test; there should be robust evidence about the association between the risk factor and serious or treatable disease; and screening should provide value for money. Therefore, we aim to answer the following questions:

Question 1: What are the reliability and validity of available methods to measure mammographic breast density?

Question 2a: Is mammographic breast density a risk factor for cancers being missed during screening (masking on mammograms/false negatives/interval cancers)?

Question 2b: Is mammographic breast density a risk factor for developing breast cancer?

Question 3: What is the test accuracy of ultrasound following mammography in comparison to mammography to detect cancer in women with dense breasts?

Question 4: For women attending breast screening in the UK, what are the cost-consequences of adding mammographic density measurements, and then ultrasound for those found to have high mammographic breast density?

Methods: Systematic reviews for each question. The search strategy combined terms for breast; screen OR screening OR “early detection of cancer”; cancer OR carcinoma OR DCIS OR malignant; ultrasound OR ultrasonography OR ultrasonics and dense OR density.

Data Sources: MEDLINE (2000-July 2017), Embase (2000-July 2017), the Cochrane Library (Cochrane Database of Systematic Reviews, CENTRAL, DARE and HTA databases) and Web of Science (Science Citation Index Expanded, Social Sciences Citation Index).

Study eligibility criteria: The key inclusion criteria are:

Participants: Women aged 47-73 attending breast cancer screening from the general population.

Interventions/comparators: Methods of measuring mammographic breast density (e.g. BI-RADS, Volpara, Quantra, Cumulus, ImageJ), and mammography plus ultrasound versus mammography only as a screening test for breast cancer.

***Outcomes**:* For density measurements: Test-retest and inter-reader reliability; concordance between methods. For the masking risk of density on mammograms: the proportion of women who develop interval cancers. For the association between mammographic breast density and breast cancer: the proportion of women who develop breast cancer (and different types of breast cancer, e.g. the more aggressive interval cancers) by density level. For supplemental ultrasound screening: recall, cancer detection, false positive and false negative rates. For cost-consequences: the cost per extra case detected.

Duplicate study selection and data extraction: Both study selection (using pre-specified inclusion and exclusion criteria) and data extraction (using a pre-piloted data extraction form) were carried out by two reviewers.

Study quality appraisal methods: Studies of reliability of density assessment were appraised using Quality Appraisal of Diagnostic Reliability (QAREL) criteria; for the association between mammographic breast density and breast cancer, we used the Quality in Prognostic Studies (QUIPS) criteria; and for the screening accuracy of ultrasound, we used the tool of the US Preventive Task Force (USPTF) and the Quality Assessment Tool for Diagnostic Accuracy Studies (QUADAS-2) form; and for the cost-effectiveness studies we used the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) form.

Synthesis methods: Data was analysed with a narrative synthesis

Results: Question 1: What are the reliability and validity of available methods to measure mammographic breast density? Our electronic search identified 2186 unique records, of which 123 were examined as full texts, and 31 papers were included, describing 27 studies. The density measurement methods examined were visual (percent density or BIRADS classification edition 3, 4 or 5); semi-automated (Cumulus, ImageJ or DM-Scan); or fully automated (Densitas, DM-Scan, Laboratory for Individualized Breast Radiodensity Assessment [LIBRA], Quantra, single energy x-ray absorptiometry [SXA] method or Volpara). We found no multi-centre that included representative samples of women and raters, that assessed repeat testing within the 2-year time-frame.

Test-retest reliability (the same images re-read by the same reader) was “moderate” to “almost perfect”, with estimates of 0.54-0.95 for visual methods; 0.92 for semi-automated methods and 0.85 for automated methods.

Inter-rater reliability was “fair” to “almost perfect”, varying from 0.38-0.96 for visual methods; and 0.83-0.92 for semi-automated methods.

In the largest real-world study, among women with consecutive mammograms interpreted by different radiologists (n = 34,271 women), at a median interval of 1.1 years (inter-quartile range [IQR] 1.0 to 1.3 years), 27.0% of women with dense breasts (BIRADS categories 3 or 4) at the first examination had nondense (BIRADS categories 1 or 2) breasts at the second examination, and 11.4% of women with nondense breasts at the first examination had dense breasts at the second examination.

Semi-automated and automated methods were more consistently reliable than visual methods.

Concordance between visual and automated methods was “fair” to “almost perfect” across different studies (0.28-0.86). Between different semi-automated methods, it was “almost perfect” (0.80-0.84). Between semi-automated and automated methods, it was “substantially different” to “substantial” (0.79); 46-52% of patients were assigned to the same quintiles by different methods. Between automated methods, there was “substantial agreement” (0.64); 50-66% of patients were assigned to the same quintiles. Even the fully-automated methods Volpara and Quantra, which are both individually highly reliable, were not interchangeable.

Results: Question 2: The searches identified 3794 studies through electronic databases; 261 records were examined at title and abstract stage, of which 54 were examined as full texts.

Question 2a: ***Is mammographic breast density a risk factor for cancers being missed during screening (masking on mammograms/false negatives/interval cancers)?*** We included seven studies, none at low risk of bias. Sample size ranged from 60 to 405,191. The studies were conducted in Australia, Belgium, The Netherlands and the USA. All found a reduced sensitivity of mammography and/or an increased risk of interval cancers with increasing mammographic breast density.

Question 2b*: *Is mammographic breast density a risk factor for developing breast cancer? We found five systematic reviews for this question and therefore conducted a review of reviews. The strength of the association between mammographic breast density and risk of breast cancer and the consistency of results between studies using varying methods, designs and locations suggests that mammographic breast density is an independent risk factor for breast cancer.

Results: Question 3: What is the test accuracy of ultrasound following mammography in comparison to mammography to detect cancer in women with dense breasts? Searches of electronic databases identified 4539 unique studies. 258 records were examined at title and abstract stage, of which 25 were examined as full texts. Eleven of the papers (reporting on nine studies) were subsequently included in the review. We found no good-quality studies.

Sensitivity of ultrasonography for women with dense breasts with negative mammography ranged from 44% to 100% between studies; specificity from 63% to 100%. The study with the highest sensitivity and specificity included around 35% of women outside the 50-70 year age range, so may not be generalisable to the UK screening population. Recall rates were 9.1 to 370 per 1000; only two of the ten studies providing data on recall rates had a recall rate for ultrasound below 10%, which is the standard from the quality assurance guidelines for breast cancer screening radiology from the NHS Breast Screening Programme (BSP)[1] for the prevalent screening round. The positive predictive value of recall (PPV_1; the chance of having cancer if recalled) ranged from 0.51% to 26.7%. Biopsy rates were between 7.3 and 66 per 1000. The positive predictive value of having a biopsy (PPV_2; the chance of having cancer if the woman has a biopsy) ranged from 2.33% to 80.8%. The rate of benign biopsies (false positives) ranged from 2.9 to 51 per 1000. Rates of additional cancer detection with ultrasound were 0 per 1000 to 7.1 per 1000. Rates of detection of small (<15mm) cancers ranged from 0 per 1000 to 2.8 per 1000. At least some of the cancers detected were of high grade and associated with positive lymph nodes. It is unclear whether the additional detection by supplemental ultrasound of small, node-negative, low grade cancers (which have a good prognosis) would be beneficial in terms of reduction of overall mortality or reduction in the rate of interval cancers or to what extent this represents overdiagnosis.

Results: Question 4: For women attending breast screening in the UK, what are the cost-consequences of adding density measurements, and then ultrasound for those found to have high mammographic breast density? We found four cost-effectiveness studies, of which only one was conducted in the UK. This UK study found that the current screening approach plus supplemental ultrasound offered to women with high mammographic breast density (defined using volumetric density grade [VDG3 and VDG4), plus magnetic resonance imaging (MRI) for women at high risk, does not appear to be a cost-effective alternative when compared with the current UK National Breast Screening Programme (NBSP):

- Incremental cost-effectiveness ratio (ICER) vs. No screening (3.5% benefits and costs discount rate [DR]): £30,772 per quality-adjusted life year (QALY) gained
- ICER vs. UK NBSP (3.5% benefits and costs DR): £212,947 per QALY gained
- ICER vs. No screening (1.5% benefits, 3.5% costs DR): £15,065 per QALY gained
- ICER vs. UK NBSP (1.5% benefits, 3.5% costs): £105,412 per QALY gained.

The first study in the USA reported that using costs of $250 per ultrasound and $2,400 per ultrasound-guided biopsy, the cost per breast cancer found was estimated to be $110,241. The second study in the USA used a theoretical calculation and reported that the cost-benefit of early detection of stage 1 disease results in annual capital cost savings of $22.75 per screened patient in the USA population. The third study in the USA reported that supplemental ultrasound screening for women with dense breasts undergoing screening mammography would substantially increase costs while producing relatively small benefits in terms of breast cancer deaths averted and QALYs gained. The ICER was $325,000 per QALY gained for women with heterogeneously or extremely dense breasts (biennial screening). Restricting supplemental ultrasound screening to women with extremely dense breasts the ICER was $246,000 per QALY gained (biennial screening). For annual screening the ICERs were even higher than biennial screening.

Only the UK study was designed as a cost-effectiveness analysis, and the intervention in that study included not only ultrasound screening for women with dense breasts but also MRI screening for women at high-risk, so the cost-effectiveness of the ultrasound component only cannot be properly established.

Discussion: The key question 1 is whether women are reliably categorised into dense or non-dense categories, irrespective of the reader, the method or the time interval (within the 2-year interval within which density is unlikely to change significantly for a woman). High quality studies would have low-risk of bias and should also be generalisable to our population in terms of the women (a large number of representative women from a general screening population) and the readers (a large number of readers within a multi-centre study of general screening, rather than single centre studies or readers specially trained for a research study). It should be noted that our review included general screening populations (which could include the usual proportion of both high- and average- risk women) and we excluded studies solely in high-risk women. In our review, the reported reliability within and between readers and the concordance between different density measurement methods varied, with many women being classified differently between readings.

Given that mammographic breast density is a risk factor for development of breast cancer (question 2b), and that breast cancer may be missed by mammography in women with dense breasts (question 2a), women with dense breasts may require supplementary screening over and above the mammography offered to women without this risk factor. For this to be feasible, it would require a) a reliable method of mammographic breast density assessment with a standardised definition of high mammographic breast density (question 1) and b) a supplementary test that was sensitive, specific, accurate (question 3) and cost-effective (question 4). Cost-effectiveness studies from the USA and the UK concluded that supplementary ultrasound was not cost-effective.

Are NSC screening criteria met?

NSC criterion 1: Questions 2a and 2b: There should be robust evidence about the association between the risk or disease marker and serious or treatable disease: **Met**. There was a strong consistent association between mammographic breast density and risk of breast cancer. There were consistent findings of reduced sensitivity of mammography and/or increased risk of interval cancers with increasing mammographic breast density.

NSC criterion 4: ***Questions 1 and 3:*** There should be a simple, safe, precise and validated screening test: **Not met**. It is difficult to validate the density methods when there is no gold standard applicable to breast density measurement. Ultrasound is not precise because it leads to large numbers of false positives, and while it can detect additional cancers not found on mammography, we do not have evidence as to whether this reduces either interval cancers or mortality, or to what extent identification of additional cancers represents overdiagnosis.

NSC criterion 14: Question 4. The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (value for money). Assessment against this criterion should have regard to evidence from cost-benefit and/or cost-effectiveness analyses and have regard to the effective use of available resource: **Not met.** There is insufficient evidence for cost-effectiveness of supplemental ultrasound, and the available evidence suggests that it is not currently cost-effective.

Strengths and limitations:

We conducted a systematic review for each of the key questions. We searched four databases, date limits were applied, and only articles in the English language were included; therefore it is possible that relevant articles might have been missed, although search terms were broad. We included a wide scope of questions including cost-effectiveness. We built on a recent review of the relevant literature and used a systematic approach to the design of our search strategies and to inclusion and exclusion and quality assessment. Sifting and data extraction were performed by two reviewers. We performed thorough quality appraisal in duplicate; no studies were excluded on grounds of quality.

A limitation of the quality assessment tool used for the studies in question 1 is that five of the eleven questions relate to blinding, with studies marked down for a lack of blinding, which may be important for research studies, but in real-world screening practice, readers would not be blinded to

previous assessment of density or clinical information, and therefore real-world studies would be inappropriately graded as lower quality. Another limitation of research studies may be their design for readers to focus all their attention on breast density, making density the most important finding on the mammograms, which is not the case in real practice in which density is usually a secondary focus of attention. Therefore, studies from real-world practice may be more informative than those in density-focused research settings.

None of the studies we found for question 2a were at low risk of bias. For question 2b, the most recent systematic review included Asian women only; the previous one contained very limited information on systematic review methods so scored poorly on the "A measurement tool to assess systematic reviews" (AMSTAR) quality criteria; the previous two focused on cancer type (Human epidermal growth factor receptor type 2 [HER2] over-expression and estrogen receptor positivity); and the earliest included review did not report the population covered or other details of the included or excluded studies.

For question 3, we updated the 2016 United States Preventive Task Force (USPTF) review, using similar search terms and quality assessment tools. However, full details of these methods were not available so relied on interpretation of the information that was present in the report. We complemented this method by carrying out our own quality assessment using the QUADAS-2 tool on both our update papers and also the original papers included in the USPTF review. However, it should be noted that some of the papers included in the USPTF review did not match our inclusion criteria (e.g. they included film mammography as well as digital). There were no good-quality studies in the question 3 update to the USPTF review – the authors of that review also noted the poor quality of the evidence base.

For question 4, four studies were included but only the one UK study was designed as a cost-effectiveness analysis, and collected and reported the required information for an economic evaluation. However, the intervention in that study included not only ultrasound screening for women with dense breasts but also MRI screening for women at high-risk, so the cost-effectiveness of the ultrasound component only cannot be properly established.

Conclusions and implications of key findings: There is strong and consistent evidence both that dense breasts increase the risk of breast cancer and decreases the sensitivity of mammography to detect cancers. Supplemental ultrasound can detect additional cancers in women with negative mammography and dense breasts, but at a cost of additional false-positives, causing anxiety for many women, unnecessary biopsies and a cost per QALY gained outside acceptable thresholds. Supplemental ultrasound in all women with heterogeneously or extremely dense breasts does not appear to be cost-effective. Focusing only on women with extremely dense breasts would be more cost-effective than including women with heterogeneously dense breasts. However, there is variation in density assessment within and between readers for visual assessment methods. Objective automated methods are more reliable than visual measures.

The implications for research include the need for:

- Assessment of methods of measuring mammographic breast density which offer consistency, reliability and validity within a general screening population, which have a

proven strong relationship to both risk of cancer and risk of masking and which are practical in terms of scale up into the screening programme.

- Stronger evidence for benefits in terms of reduction in interval cancers or breast cancer mortality from supplemental ultrasound after mammographic breast density assessment.
- A randomised controlled trial including cost-effectiveness assessment to provide the necessary answers to the question of whether density assessment followed by ultrasound for women with dense breasts would be clinically and cost-effective within the screening programme. Follow up long enough to assess the different types of cancer found, along with any reductions in interval cancers, would be required in order to address the issue of potential overdiagnosis.

The implications for practice

If density assessment followed by supplementary ultrasound screening were undertaken in the current NHS breast screening programme, women could be categorised differently between readers or screening occasions unless a standardized programme-wide method of density assessment were used. Such a programme however would be likely to lead to increased anxiety and resource use (for women identified as at higher risk who might not actually be at higher risk), and to confusion for women whose categorization changed. Our review suggests that the numbers of false positives and additional biopsies are unlikely to be justified and that there is as yet no clear cost effectiveness evidence to balance the benefits, harms and costs.

Systematic review registration number: CRD42017081213

Source of funding: PHE Screening

Plain text summary

Breasts are made up of a mixture of fibrous and glandular tissue and fatty tissue. Breasts are considered dense if they have a lot of fibrous or glandular tissue but not much fat. Having dense breast tissue may increase the risk of getting breast cancer. Dense breasts also make it more difficult to spot cancer on mammograms. Dense tissue appears white on a mammogram. Lumps, both benign and cancerous, also appear white. So mammograms can be less accurate in women with dense breasts. Studies have shown that ultrasound can help find breast cancers that can't be seen on a mammogram. However, ultrasound shows up more findings that are not cancer, which can mean testing and biopsies that aren't needed. Breast density is read on a mammogram by a radiologist, or using automated methods. We wanted to answer the question of whether measurement of mammographic breast density is reliable, that is, will the same reader (at different times), or different readers, or different measurement methods, always give the same answer about whether breasts are dense or not? This is important to find out if it is worthwhile measuring mammographic breast density, and doing extra tests (ultrasound) on women with dense breasts.

We carried out a systematic review of the literature to find information about the reliability of different mammographic breast density measurement methods among women attending breast cancer screening. We found that reliability varied between the studies. For example, in the largest study, among women with two mammograms interpreted by different radiologists, around a third had a different density assessment at the 2 examinations. With density described in two categories (dense or nondense), nearly a fifth of women had different density ratings at the 2 examinations; around a quarter of women with dense breasts at the first examination were stated to have nondense breasts at the second examination, and around a tenth of women with nondense breasts at the first examination were stated to have dense breasts at the second examination. Readers vary in their interpretation of mammographic breast density: some readers rated less than a third of women with dense breasts, while other readers rate over half of women with dense breasts. There was a lot of variation in density assessment within and between readers in the studies we found. The automated methods appear to be more reliable than human readers, but so far there isn't enough high-quality evidence to support this, and even automated methods do not give the same answers as each other, as they define density differently.

We found several systematic reviews suggesting that women with dense breasts are more likely to develop breast cancer, and other studies reporting that mammograms are less likely to pick up cancers if women have dense breasts.

We updated a recent large USA review of ultrasound following a negative mammography screen, and found that it still missed some cancers, while flagging up many areas of concern that turned out to be false alarms. We concluded that until more reliable methods of measuring mammographic breast density are available, there is not enough evidence to support supplemental ultrasound screening for women based on mammographic breast density measures in routine screening practice.

We found four studies giving information on cost-effectiveness of additional ultrasound in women with dense breasts. The extra ultrasounds substantially increased costs while finding relatively few

extra cancers, while causing many women anxiety because of “false-positive” tests (when concern over the scan results meant women had to have unnecessary biopsies which turned out not to be cancers). Overall the addition of ultrasound did not appear to be cost-effective.

Section 1: Introduction

1.1 Background

Breast cancer is the most common cancer in the UK, for example, there were 55,200 new cases in 2014, almost all in women[2]. The risk varies with factors such as age, age at menarche, parity, age at birth of first child, age at menopause, body mass index (BMI), first-degree relatives with breast cancer, use of hormone replacement therapy (HRT) and breast density (the proportion of fibroglandular tissue in the breast).[3,4] Around a third of female invasive breast cancer cases in England are detected by screening,[5] another third occur in the interval between mammograms,[6] and the rest are found in women outside the screening age range, or in men.

Mammograms are offered every 3 years in the UK National Health Service Breast Screening Programme (NHSBSP).[6] Interval cancers have a worse prognosis than screen-detected cancers, so identifying women at higher risk of interval cancers (e.g. women with dense breasts) and offering them tailored screening interventions may improve the effectiveness of the NHSBSP.[6,7] A recent report from the Public Health England (PHE) Working Party for Higher Risk Breast Screening suggests that if a specific programme for screening women with high risk becomes a priority, a way of identifying them will be needed, e.g. by detection of high density on a mammogram.[8]

There are several methods for measuring density in mammography.[9] These include visual methods (assessment of the mammogram by a reader), semi-automated methods (the reader uses a computer-assisted technique) or fully automated methods (density assessed by a computer algorithm). However, there is no gold standard measurement of mammographic breast density applicable to all breast density measurements, and different measurement methods define the concept in various ways, limiting the concordance between methods. While MRI has been suggested as a type of gold standard, discrepancies occur between breast density measurement methods and this gold standard, particularly at higher densities.[10]

Visual assessments: the reader estimates the breast area, absolute density, absolute non-density/fat (all in cm^2) and percent density (the ratio of parenchyma to fat as seen on mammography), from mammographic images (i.e. an area-based method). Methods include:

The four categories of mammographic breast density defined by the American College of Radiology's *Breast Imaging Reporting and Data System (BI-RADS) 4th edition* criteria:[11]

- The breasts are almost entirely fatty (percent density <25%)
- There are scattered areas of fibroglandular density (percent density 25–50%)
- The breasts are heterogeneously dense, which may obscure small masses (percent density 51–75%)
- The breasts are extremely dense, which lowers the sensitivity of mammography (percent density >75%).

In 2013, the *BI-RADS guidelines (fifth edition)* changed.[12] Categories A, B, C, and D are (a) fatty, (b) scattered density, (c) heterogeneously dense, and (d) extremely dense, but the percentages were removed, and more emphasis was given to the potential masking of the dense tissues.[12,13] In the new guidelines, a breast could still be classified as dense even if it is < 50% glandular but the radiologist is concerned about an area of dense tissue that could potentially mask an underlying cancer.[12] Removing the percentages from the density assessment guidelines might be expected to

result in a reader's observation becoming more subjective, with an associated drop in intra- and inter-reader agreements, and an increase in the proportion of women categorised as having dense breasts and therefore becoming candidates for supplemental screening; both of these effects were apparent in a study comparing the BIRADS 4th and 5th editions.[12]

Semi-automated methods include:

Cumulus, *QWIN* and *DM-Scan*

In these methods, the operator outlines the total breast and sets a threshold to separate the dense tissue from the fatty tissue, so density is calculated as the dense area divided by the total breast area.[9,14-16]

Fully automated methods include:

Area-based methods:

- The fully-automated version of *DM-Scan*, in which supervised pixel labelling is used to train a fully-automated classifier.[15]
- *Densitas' DM-Density* calculates the percentage of the breast image composed of dense tissue, accounting for its texture and distribution, in the "for presentation" digital image.
- The area-based *ImageJ*-based method, a fully-automated approach mimicking Cumulus by measuring several image parameters and choosing those shown to predict Cumulus density in a training set of images with known Cumulus-density readings.[9] The selected parameters are then used in a regression model to estimate percent density values in other images.
- The *Laboratory for Individualized Breast Radiodensity Assessment (LIBRA)*, which generates area-based measurements of breast area, dense tissue area and percentage density.[17] The algorithm first identifies and extracts the breast region, then segments the dense tissue within the breast by using a combination of fuzzy c-means clustering and support vector machine classification.

Volume-based methods:

- *Volpara* is a volumetric method (i.e. estimated breast, absolute dense and absolute non-dense volumes [all in cm^3] and percent density, from digital images) using an algorithm to assess the x-ray attenuation of tissue between the image detector and the x-ray source on the basis of the pixel values on the images.[18] Percent volumetric mammographic breast density is calculated as the ratio of fibroglandular tissue volume to total breast volume. This quantitative volumetric breast density value is mapped to an automated density grade using preset thresholds (automated density grade 1: <4.5%; grade 2: ≥4.5% and <7.5%; grade 3: ≥7.5% and <15.5%; grade 4: ≥15.5%) to map onto the BIRADS categories. It averages estimates from craniocaudal (CC) and mediolateral oblique (MLO) views for each breast and has an outlier removal process, and uses physical modelling of mammographic systems. Volumetric breast density measurement is based on the physical composition of the breast, compressed breast thickness, and x-ray information (tube potential [kVp], tube current [mAs], filter type and thickness).
- *Quantra* averages estimates from craniocaudal (CC) and mediolateral oblique (MLO) views for each breast using physical modelling of mammographic systems to calculate volumetric breast density (dense tissue volume/total breast volume) and area percentage breast

density (area of fibroglandular tissue/total breast area).[19] Volumetric breast density measurement is based on the physical composition of the breast, compressed breast thickness, and x-ray information (tube potential [kVp], tube current [mAs], filter type and thickness). Quantra segments the estimated volumetric breast density to generate fractional quantised breast density (q_abd) values for each mammographic view. These are averaged to a total Q_abd for each patient (rounded) so Q_abd 1 is ≤1.44; Q_abd 2 is 1.45 to 2.44; 3 is 2.45 to 3.44; 4 is ≥3.45. Quantra Q_abd values 1 to 4 then map onto BIRADS 1 to 4 categories.

- *Single energy x-ray absorptiometry (SXA)* uses a calibration phantom (made from materials that mimic glandular-fatty tissue ratios) on the unused corner of the compression paddle of the x-ray machine; it can only process CC images.[9] An algorithm then analyses the digital image and estimates breast thickness and amount of fibroglandular density at each pixel. The pixel-specific estimates are then summed up to produce total breast estimates for dense tissue volume (in cm^3), and volumetric percent density.

Of note, methods for research purposes only include Cumulus, ImageJ, LIBRA and SXA, while commercially-available methods include Densitas, Quantra and Volpara.[20]

In one UK study (n=1969), the performance of three area-based approaches (BI-RADS, the semi-automated Cumulus, and the fully-automated ImageJ-based approach) and three fully-automated volumetric methods (Volpara, Quantra and SXA) were assessed in full-field digital mammography (FFDM) images from cases (the unaffected breast of women with newly-diagnosed breast cancer) and controls (women without breast cancer).[9] For all methods, percent density was lower with increasing age, BMI, parity, postmenopausal status, and cancer risk was higher with higher density.[9] However, the discrimination between cases and controls by density was low for all methods, highlighting its limited value in individual risk prediction.[9] Practical issues identified in the study were:

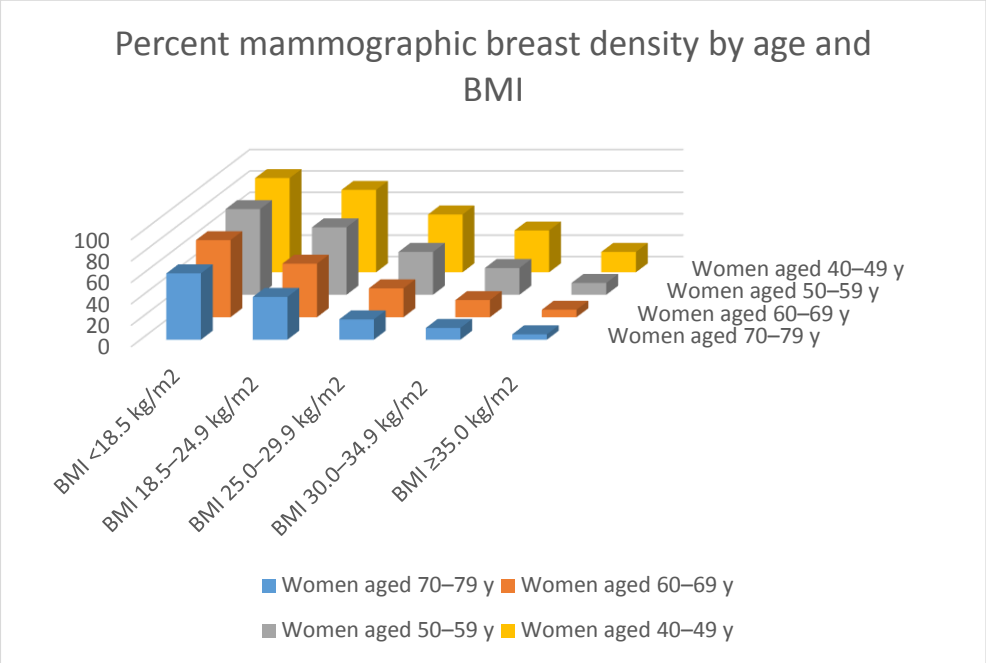
- The methods were based on raw ("for processing") images, which need to be saved. Currently, only processed ("for presentation") images are routinely saved in most screening/clinical settings.[9]
- SXA readings were missing for many participants due to lack of a phantom, limiting its use in busy clinical settings, and it cannot be applied retrospectively to historical images.[9]
- Quantra (version 1.3) produced a digital image with the density measurements superimposed on it, which is convenient in screening/clinical settings, but not efficient in large-scale studies as the density measurements for analysis would have to be extracted manually. Different versions of Volpara (clinical and research) are available. There are currently no stand-alone software packages for SXA or ImageJ, limiting widespread implementation.[9]
- The volumetric methods attempted to estimate volumetric density from two-dimensional images, supplemented by information on the third dimension (using phantoms, breast thickness, or plate tilting). Three-dimensional imaging techniques, e.g. tomosynthesis or MRI, are not widely used clinically.[9]

Visual density assessment methods show a strong relationship between density and breast cancer, despite inter-observer variability, but are impractical for population-based screening.[21] Cumulus was developed to improve reproducibility but also requires trained observers, and although separating the breast from the mammogram background is reproducible, assessment of the best threshold to

separate dense tissue from fat is less reproducible.[21] Automated methods may be more practical for risk stratification.[21]

It is of note that breast density varies over time, with age, BMI and menopausal status. For example, in a USA mammography study[22] (including 216,783 screening mammograms from 145,123 women), the percentage of mammograms reported as showing dense breasts varied by age and BMI as shown in the following Figure: 1

Figure 1. Percent mammographic breast density by age and BMI



Similar reductions in density with age are broadly seen in various ethnic groups including Black, Eastern Mediterranean, East Asian, South Asian/Malay, Mestizo/Hawaiian and White women, although absolute values of percent density vary.[23]

Conventional film mammography screening is known to reduce breast cancer mortality among women aged 50–69 years, but mammography has lower sensitivity in younger women, partly due to their greater breast density.[24] Digital mammography is now standard throughout the UK,[21] so it is important to assess methods of density assessment on digital mammograms for risk assessment, which could be used to inform interventions (e.g. weight loss for overweight/obese women) and/or supplemental screening methods in women found to be at increased risk.

1.2 Rationale, objectives and key questions

In the current UK breast screening pathway (see Figure 2), women in the general population aged 50–70 years receive mammography testing every 3 years with no density measurement, and no ultrasound (except as part of the follow-up tests for screen positives). Mammography screening takes 6 minutes to perform and results are returned within two weeks after examination by two independent experts (radiologist, radiography advanced practitioner or breast clinician); disagreements may be resolved by consensus or arbitration involving another reader. The potential pathway under investigation includes the addition of breast density estimated from mammograms (either every screen or less frequently) (see Figures 3 and 4). The aim of which would be to identify

women with a risk higher than the general population, based on mammographic breast density, who might benefit from an enhanced screening programme (using ultrasound). Women with dense breasts could then be offered ultrasound in addition to mammography at screening. Ultrasound and mammography may be at the same or different appointments (and therefore ultrasound screening may be given to all women with dense breasts [if the mammogram outcome is not yet known; Figure 3] or only be given to mammography-negative women [if only mammogram-negative women are recalled for ultrasound after the mammogram has been read; Figure 4]). Handheld ultrasound takes 20 minutes but results are available immediately; automated ultrasound is reported later. (The density measurements are also applicable to future potential changes to screening, for example digital tomography could be introduced for dense breasts only.) Women receive further investigations (e.g. biopsy for definitive diagnosis) if this is indicated by either ultrasound or mammography.

Figure 2: Current pathway

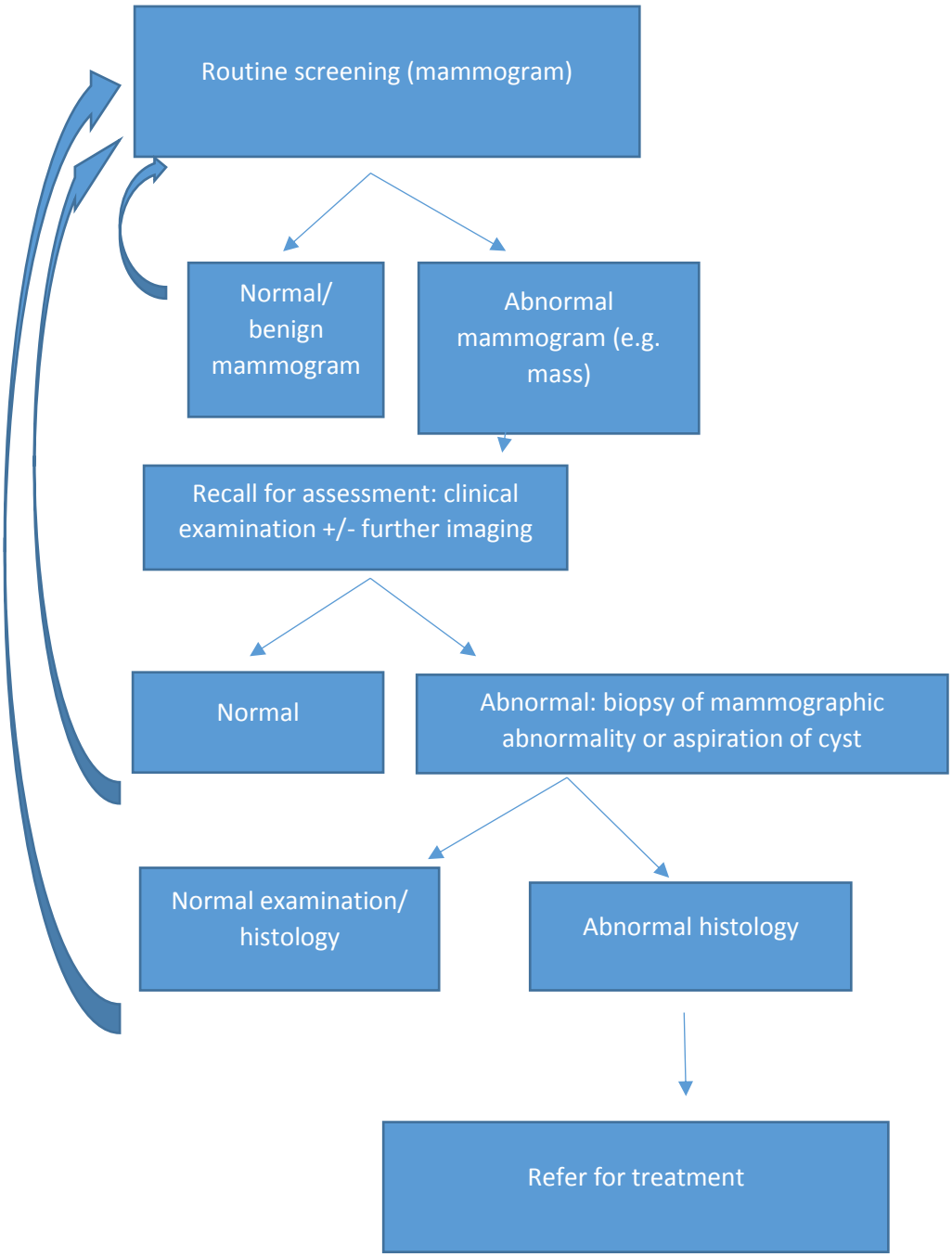


Figure 3: Pathways under investigation: all women identified with dense breasts get ultrasound

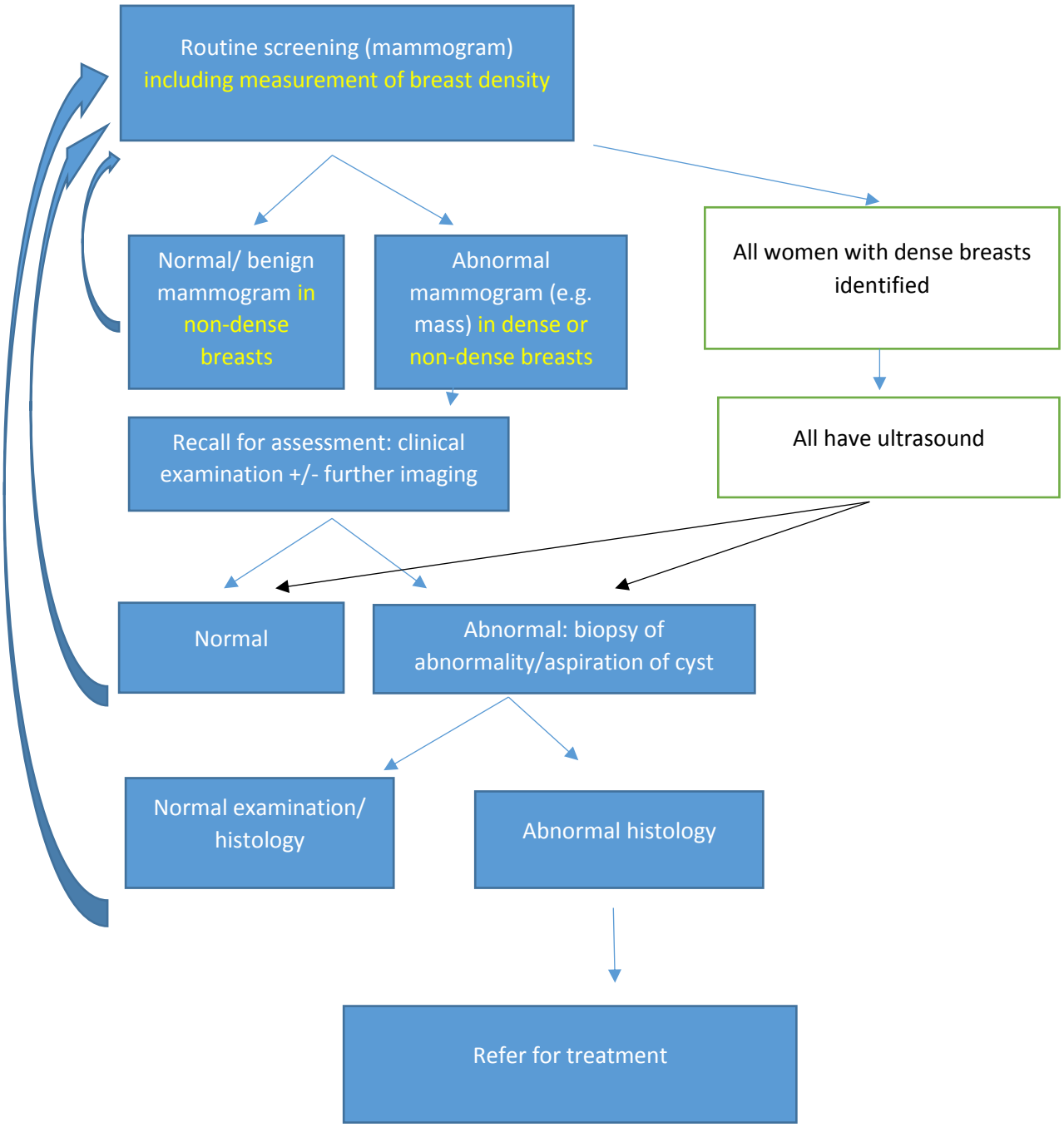
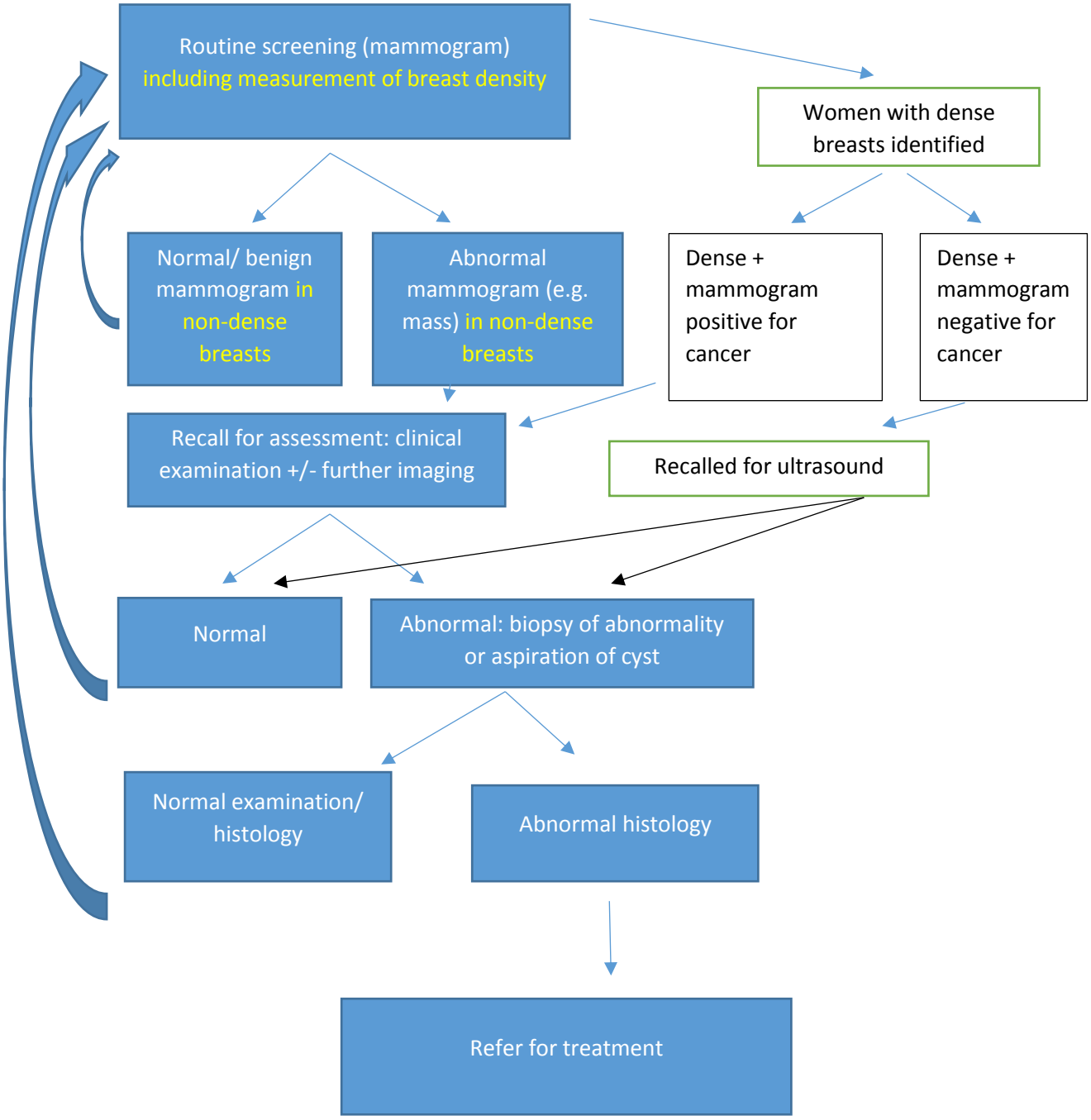


Figure 4: Pathways under investigation: women identified with dense breasts whose mammogram is negative for cancer get ultrasound



Policies about supplemental screening vary. For example, in the USA, legislation in many states requires that providers notify patients about their mammographic breast density, and in some cases, requires insurance coverage of subsequent supplemental screening.[25] This raises questions for women and their doctors about the interpretation of screening results and the need for additional testing.[25] If the assessment of mammographic breast density is not reliable (e.g. variability in breast density determinations between readers or over time), this could undermine women's confidence in the screening process and leave them uncertain about their risk for breast cancer.[25] Therefore it is important to determine the reliability of the methods of assessment of mammographic breast density.

To assess evidence about the association between mammographic breast density and serious or treatable disease, it is important to understand to what extent breast density is associated with various subtypes of breast cancer, including interval versus screen-detected cancer; invasive versus in situ lesions; and characteristics relating to the degree of differentiation, aggressiveness or receptor status of cancers. Ultrasound as an additional screening test in women found to have dense breasts could detect more cancers than mammography alone, but could also lead to increases in recall and biopsy rates, anxiety, over-diagnosis and increased costs.[25] It is therefore important to assess both the test characteristics (sensitivity, specificity, false negatives, false positives etc.) and the cost consequences of supplemental screening, plus limited resource availability, particularly in regard to the personnel and time required for image acquisition and interpretation.

In 2012, the American College of Radiology published a position statement urging strong consideration of the benefits, possible harms and unintended consequences of including breast parenchymal information in the information given to women.[26] In particular they mentioned that:

- visual assessment of breast density is not reliably reproducible;
- the significance of breast density as a risk factor for breast cancer is highly controversial, and there is no consensus that density per se confers sufficient risk to warrant supplemental screening;
- while supplemental screening can detect cancer not found via mammography, it also results in additional false positive examinations and increases the number of benign breast biopsies, and there is no randomised trial data that shows that adding ultrasound to mammography screening saves lives; and
- there are costs involved in the additional testing.[26]

It is therefore important for the UK to review the evolving evidence base and consider policy in the light of the reliability of density measurement and its significance (independent of other potential risk factors such as age, BMI, parity, family history etc.) as a risk factor for breast cancer, the properties of ultrasound as a supplemental screening test and its cost consequences.

1.3 Objectives: Evidence Review

We undertook a systematic review according to the UK NSC guidelines.[27] The UK NSC has produced criteria for appraising the viability, effectiveness and appropriateness of a screening programme[28] (see Appendix 7). The overall aim of this review was to determine the balance of benefits and harms, and the costs of measuring mammographic breast density, and of offering women with dense

breasts an ultrasound test. Table 1 below shows the four key questions of the review and how they map onto the NSC appraisal criteria.

Table 1: Key questions and NSC criteria

Key question for the review	**NSC criterion**
Question 1: What are the reliability and concordance of available methods to measure mammographic breast density?	NSC criterion 4: There should be a simple, safe, precise and validated screening test.
Question 2: 2a: Is mammographic breast density a risk factor for cancers being missed during screening (masking on mammograms/false negatives/interval cancers)? 2b: Is mammographic breast density a risk factor for developing breast cancer?	NSC criterion 1: There should be robust evidence about the association between the risk or disease marker and serious or treatable disease.
Question 3: What is the test accuracy of ultrasound following mammography in comparison to mammography to detect cancer in women with dense breasts?	NSC criterion 4: There should be a simple, safe, precise and validated screening test.
Question 4: For women attending breast screening in the UK, what are the cost-consequences of adding density measurements, and then ultrasound for those found to have high mammographic breast density?	NSC criterion 14. The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (value for money). Assessment against this criteria should have regard to evidence from cost benefit and/or cost effectiveness analyses and have regard to the effective use of available resource.

Section 2: Methods

2.1 Methods of developing the protocol

We undertook a systematic review according to the UK NSC's requirements. We incorporated guidance from commissioners and experts. The protocol is registered at PROSPERO: the International Prospective Register of Systematic Reviews (registration number: CRD42017081213).

2.2 Identification and selection of studies

Separate searches were conducted for each of the key questions, and the results downloaded into Endnote and de-duplicated. Full details of the searches are provided in Appendix 1. The search strategy comprised searching of electronic bibliographic databases, contact with experts in the field, and scrutiny of the references of included studies and relevant systematic reviews. We searched the following electronic databases: MEDLINE (2000-July 2017), Embase (2000-July 2017), the Cochrane Library (Cochrane Database of Systematic Reviews, CENTRAL, DARE and HTA databases), and Web of Science. The search was initially from 1 January 2000 for Q1 and Q2 and from 1 January 2005 for Q3 and Q4. However, it was planned that if recent a single high quality systematic review was identified that answered the research question, we would carry out an update of that existing systematic review including eligible studies published subsequent to the search date for the systematic review, to avoid duplication. If several systematic reviews were available for a question, we would conduct an overview of reviews for that question. The inclusion and exclusion criteria for each of the key questions are shown in Table 2.

Papers (non-systematic reviews) reporting pooled analysis from multiple studies, i.e. the studies had different sites/inclusion criteria but were not selected by a systematic search, were reference checked to ensure that eligible studies within the pooled analysis were included as individual studies in our review. Papers reporting studies conducted by the same organisation (same inclusion criteria/protocol) but different years/cohorts/sites were treated as a single study for data extraction. A paper reporting two separate cohorts (analysed separately) was treated as two separate studies. Multiple publications from the same study/cohort were data extracted together to avoid double counting. The most appropriate analyses were selected as the main findings (e.g. involving the largest number of women).

Table 2. Inclusion and exclusion criteria for the four key questions

Key question	Inclusion criteria						Exclusion criteria
	Population	Intervention / Index test	Reference standard / comparator	Outcomes	Study design	Type and language	
1. What are the reliability and concordance of available methods to measure mammographic breast density?	Women aged 47-73 attending breast cancer screening from the general population	Using digital mammograms only (not film): BI-RADS scale scored by a single qualified reader BI-RADS scale scored by a group consensus of readers Volpara Quantra Densitas LIBRA Cumulus Madena ImageJ (Stratus) Single energy x-ray absorptiometry (SXA) DM-Scan Left breast/right breast comparison The Royal Australian and New Zealand College of Radiologists (RANZCR)	As for index test	Test-retest reliability Inter-reader reliability Concordance between methods Positive and negative concordance between pairs of tests Comparison of characteristics of discordant cases: in particular comparison of risk of breast cancer and measures of missing cancers at screening such as interval cancers.	Cross-sectional studies, test quality studies nested within RCTs or cohort studies, case-control studies, and test sets involving multiple blinded readings of mammography Minimum number of participants = 100	English language Full text report From 2000 onwards	**Population outside scope:** **Age**: Studies in which ALL the women fall OUTSIDE the age range 47-73 years. **Population outside scope**: high risk population e.g. women with clinically significant Breast Cancer (BRCA) 1/2 mutations or other familial breast cancer syndromes or women with previous breast cancer; symptomatic women, i.e. diagnostic (rather than screening) mammograms. Papers with mixed screening/diagnostic populations were excluded (unless screening populations were reported separately). Other: e.g. studies on mastectomy or post-mortem specimens/rare tumours (e.g. malignant phyllodes)/

2a: Is mammographic breast density a risk factor for cancers being missed during screening (masking on mammograms/ false negatives/ interval cancers)? **2b: Is mammographic breast density a risk factor for developing breast cancer?**	Women aged 47-73 attending breast cancer screening from the general population	Using digital mammograms only (not film): BI-RADS scale scored by a single qualified reader BI-RADS scale scored by a group consensus of readers Volpara Quantra Densitas LIBRA Cumulus Madena ImageJ (Stratus) Single energy x-ray absorptiometry (SXA) DM-Scan RANZCR	As for index test	Single or head to head studies (1 or more types of test): Proportion of women who have an interval cancer after screening by density for each test Proportion of women who have breast cancer by density for each test (includes reporting of absolute risk which is of particular interest in low density groups) Distribution of cancer type by risk group for each test Odds ratios (OR) or risk ratios (RR) from unadjusted univariable	Head to head or single arm studies: RCTs, prospective cohort, case-control, nested case-control, or cross-sectional studies	English language Full text report From 2000 onwards	animal/phantom/simulation studies. **Intervention/comparator outside scope**: studies assessing one density measure (e.g. Volpara) assessing two views (CC/MLO) were not included as test-retest samples for reliability; **a**ssessing density of a mass rather than of the breast as a whole; CT; MRI. Studies of cancer risk models were not included for question 2 unless they reported the association between density and cancer risk (unadjusted or age-adjusted) separately from other factors in the risk model (although multivariate analyses were also extracted). **Outcome outside scope**: e.g. molecular or genome studies/ pre-operative assessment of tumour size/ breast density as an outcome of intervention studies/ studies detecting change in density over time >2 years or before versus after the menopause. **Study design outside scope:** e.g. Survey/case report/grey

				models of density as a predictor of risk (and models adjusted for age only). Results to be stratified by age: <40 / 40-49 / 50-70 / >70; or <46 / 47-73 / >73 years			literature (i.e. editorials, letters, commentaries and conference abstracts). **Other not relevant:** e.g. different topic.
3. What is the test accuracy of ultrasound following mammography in comparison to mammography to detect cancer in women with dense breasts?	Women aged 47-73 with dense breasts attending screening from the general population	Ultrasound (automated/tomography [in the mammography machine or as a separate machine], or handheld if the whole breast is assessed) as a screening test for breast cancer Mammography (digital not film) as a screening test for breast cancer	Biopsy test for cancer, and follow up to interval cancers	For cancer detection: Sensitivity and specificity Positive and negative predictive values 2x2 tables. Characteristics of extra cancers detected by US only and mammography only (comparison of discordant cases or	Head to head (mammography versus mammography plus ultrasound) test accuracy studies in the same population, or test accuracy of ultrasound in a mammography-negative population; cohort studies; randomised controlled trials	English language Full text report From 2005 onwards (cut off for relevant ultrasound technology)	

				diagonal cells in 2x2 table) a) invasive cancers only; b) Ductal carcinoma in situ (DCIS) separately where reported; c) both invasive + DCIS (total cancers). % DCIS Prognosis measures, grade, stage, nodal involvement Tumour type (lobular or ductal) estrogen receptor (ER)/ progesterone receptor (PR) status Size. Risk of overdiagnosis (especially with repeated			

				measurement of breast density)			
4. For women attending breast screening in the UK, what are the cost-consequences of adding density measurements, and then ultrasound for those found to have high mammographic breast density?	Women aged 47-73 invited to mammography screening from the general population	Supplemental ultrasound	Mammography only	Cost per extra case detected Cost per extra case detected by type (e.g. cost per extra high risk case detected invasive? Nodes involved?)	Cost consequence model, or simple addition of costs in particular cost of density measurements and cost of ultrasound; cohort studies; randomised controlled trials; systematic review of these study designs	English language Full text report From 2005 onwards (cut off for relevant ultrasound technology)	

2.3 Study selection

Firstly, we assessed any systematic reviews for each question of this review. The titles and abstracts of articles from the searches were assessed independently by two reviewers (see Table 2 for inclusion/exclusion criteria). Disagreements about inclusion/exclusion were resolved by retrieval of the full publication and consensus agreement. Full copies of all studies deemed potentially relevant were obtained and assessed independently by two reviewers. Any disagreements were resolved by consensus or discussion with a third reviewer. Details of studies excluded at each stage were documented (see Appendix 3).

2.4 Data extraction

Data were extracted by a single reviewer using a piloted data extraction sheet. All of the extracted data were checked by a second reviewer. Any disagreements were resolved by consensus or discussion with a third reviewer. An example data extraction sheet is provided in Appendix 4.

2.5 Assessment of quality/risk of bias in individual studies

Papers for question 1 were assessed using the Quality Appraisal of Diagnostic Reliability (QAREL) Checklist.[29] Papers for question 2a were assessed using the Quality in Prognostic Studies (QUIPS)[30] and systematic reviews for question 2b were assessed using the AMSTAR criteria.[31] Papers for question 3 were planned to be assessed using the modified quality assessment tool for diagnostic accuracy studies (QUADAS-2);[32] however, a high-quality systematic review was identified (USPTF)[25] and updated. Therefore we used the same quality assessment criteria as that review (USPTF criteria), in addition to the QUADAS-2 as originally planned. For question 4, papers were assessed using the Consolidated Health Economic Evaluation Reporting Standards (CHEERS) checklist.[33]

Quality appraisal was undertaken independently by two reviewers, with disagreements resolved through consensus or in discussion with a third reviewer. Quality assessment forms are shown in Appendix 5.

2.6 Evidence synthesis methods

Results of each question were narratively synthesised. Where outcomes of interest were not reported, we calculated values where sufficient data were reported. For question 1, kappas were interpreted as follows: 0.01–0.20 represent slight agreement, values of 0.21–0.40 represent fair agreement, those between 0.41–0.60 represent moderate agreement, values between 0.61–0.80 represent substantial agreement and values between 0.81–0.99 represent almost perfect agreement.[34] The intra-class correlation coefficient (ICC) is equivalent to the weighted kappa. ICC of less than 0.40 represents poor agreement, 0.40-0.59 represents fair agreement, 0.60-0.74 represents good agreement, and 0.75-1.00 represents excellent agreement.[35] For question 3, sensitivity and specificity of ultrasound in women with dense breasts and negative mammography were analysed using a Forest plot.

Section 3: Results

3.1 Key question 1 (reliability and concordance)

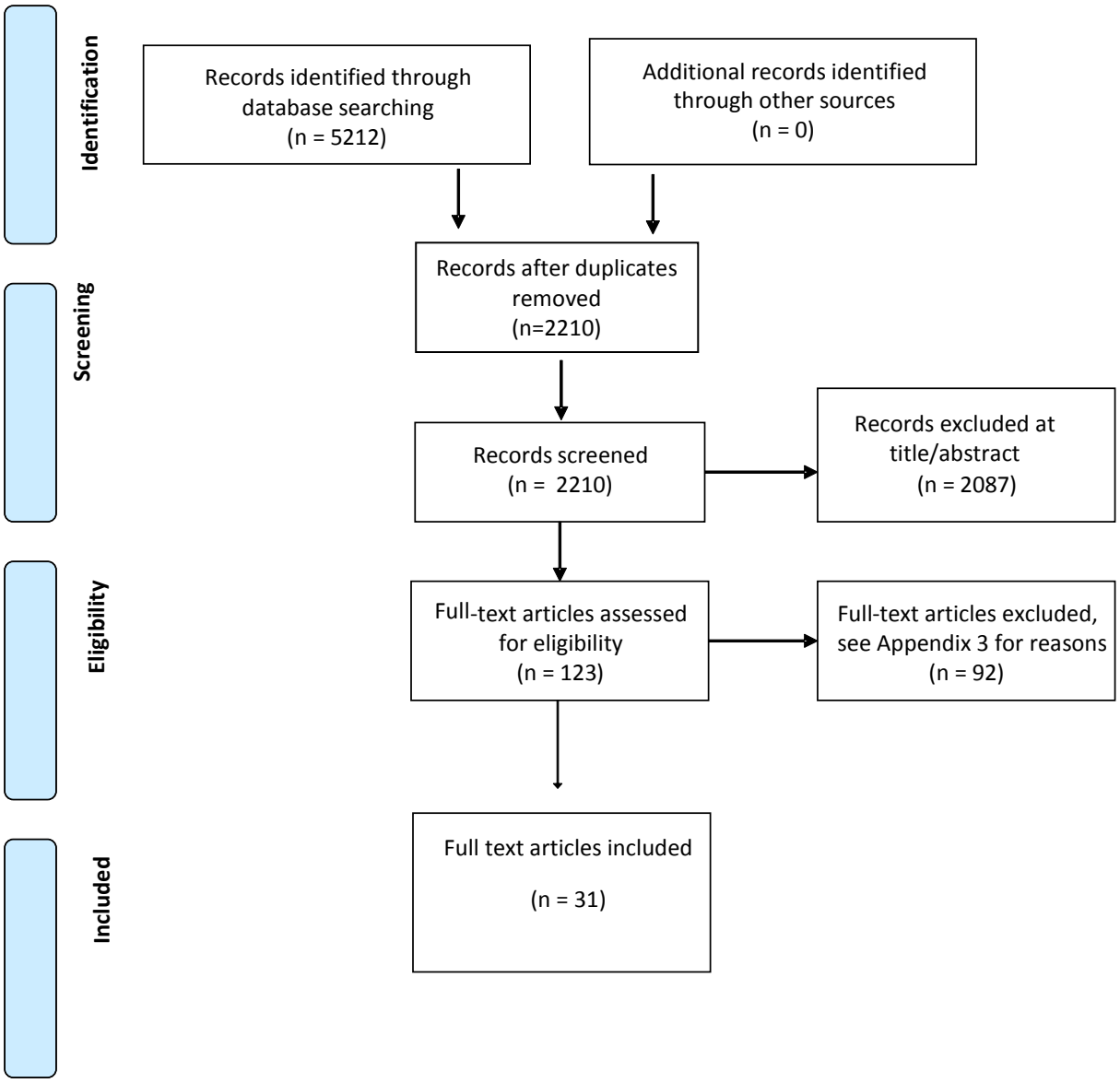
What are the reliability and concordance of available methods to measure mammographic breast density?

This relates to NSC criterion 4: “There should be a simple, safe, precise and validated screening test.”

3.1.1 Description of the evidence

Figure 5 provides the PRISMA flow diagram for the reliability and concordance question. Our electronic search identified 2186 unique records, with no additional records identified through other sources. One hundred and twenty-three were examined as full texts. Ninety-two studies were excluded at full text stage; these are listed with the reason for exclusion in Appendix 3. This left 31 papers, reporting on 27 studies, which were included in the review.

Figure 5: PRISMA flow chart for question 1



3.1.2 Characteristics of included studies

Thirty-one papers, reporting on 27 studies, were included, which are summarised in Table 3 and Appendix 6 (Question 1 Tables a and b). Sample sizes ranged from 100 to 145,123 women. The studies were conducted in Australia,[19,36] Canada,[37] India,[38] Israel,[39] the Netherlands,[13,16,40] Norway,[41] the Republic of Korea,[42-45] Spain,[15,46] Sweden,[47] the UK[9] and the USA.[12,14,17,18,22,48-52] The approach to density measurement, and the type of images used, varied between studies, with some studies using more than one method. Visual density measurement methods (percent density[37] or BIRADS classification edition 3,[15,49] 4 [9,12,16,18,19,22,40-42,44,46,47,50,51] or 5,[12,13,17,36,38,39,45,48,51,52] or version not stated[14]) were assessed in 25 studies.[12-19,22,36-42,44-52] Semi-automated methods (Cumulus,[9,14,15,43] ImageJ[9] or DM-Scan[15]) using processed images were assessed in four studies.[9,14,15,43] Fully automated methods (Densitas,[37] DM-Scan,[15] LIBRA,[17] Quantra,[9,13,19,41] SXA[9] or Volpara[9,13,14,18,38,40,42,44,47,50,52]) were assessed in raw ("for processing") images,[9,13,14,19,40,44,47,50,52] processed ("for presentation") images,[15,17,37,42] mixed raw and processed images,[18] and in three studies the image type was not stated.[38,41,45] For the inter-rater reliability studies, the number of raters ranged from two[16,37,38,43] to eighty-three,[22] and for the test-retest studies, the time between ratings ranged from 1 day[46] to 30 months.[40] Concordance between measures was examined in 17 studies.[9,13-15,17-19,37,38,40-42,44,45,47,50,52]

Table 3: Methods, quality summary and limitations of included studies in question 1

Study	Population (n)	Interventions/ Comparator	Outcome	No. centres; country	Quality: QAREL criteria met/not met/ unclear/ not applicable out of total 11 domains	Sample repress-entative?	Readers repress-entative?	Time <2 years between tests?	Limitations
Abdolell 2013[37]	Digital mammograms – no further information (n=138)	Densitas and visual percent density assessment	Inter-rater reliability; concordance between Densitas and visual assessment	1; Canada	3/0/7/1	Unclear	Yes	Unclear	The Pearson correlation coefficient (ρ) provides an inadequate, inflated, and overoptimistic measure of the level of agreement. This measure is not eligible for our review.
Alshafeiy 2017[48]	Consecutive women undergoing screening with digital 2D mammography and tomosynthesis with a negative or benign (category 1 and 2) outcome (n=309); mean (SD) age 65.7 ± 11.4 years (range, 35–93 years).	BI-RADS 5th edition from digital 2D images	Interreader agreement	1; USA	4/1/3/3	No	Yes	Yes	Relatively small number of readers from a single institution; results may differ in a larger study with more readers. No reference standard for breast density
Conant 2017[17]	Women with 2D bilateral MLO view synthetic digital mammogram (sDM) and standard dose "For	BIRADS 5th edition; LIBRA algorithm in DM	Analysis of variance to determine whether the automated percent density estimates for DM	1; USA	1/7/0/3	No	No	N/A	A single area-based density estimation method using data from a single institution

	presentation" DM images available (3668 women with 7336 MLO images)		varied significantly according to the corresponding BIRADS breast density categories						
Destounis 2017[18]	Women diagnosed with cancer within the screening programme; mean (SD) age 62.1 (11) (n=595)	BIRADS 4th edition, from previous normal mammogram vs. Volpara v1.4.2 from previous normal mammogram if raw images available or contralateral breast if raw images not available	Agreement between visual BIRADS and automated density grade	1; USA	3/1/5/2	No	Unclear	Yes	Interval cancers not differentiated between true interval, missed or mammographically occult (i.e. masked by dense tissue).
Ekpo 2016.[36]	Women who underwent digital breast tomosynthesis (DBT) investigation in 2015 and had a prior DM obtained in 2014 (n=234)	BI-RADS 5th edition	BI-RADS 5th edition inter-reader reproducibility	1; Australia	4/1/3/3	No	Yes	Yes	The proportion of BIRADS D density category in the dataset is higher than that of a typical population distribution, as women that have DBT subsequent to DM are more likely to have dense breast than fatty breasts. No agreed standard for BD assessment.
Ekpo 2016.[19]	Females who underwent screening mammography between March and July 2014 (n=292)	Quantra 2.0 vs. BIRADS 4th edition	Agreement between each radiologist and the majority report. Inter-reader agreement was	1; Australia	7/1/3/0	Unclear	Yes	Yes	The high level of agreement between the 6 radiologists may be due to the readers all working in the same practice; it is possible they

			assessed by comparing the first assessment of the radiologists in pairs. Intra-reader agreement was assessed by comparing the first and second readings of each radiologist.						would demonstrate considerable inter-reader variability with readers from different practice, limiting generalizability. Using the majority report in Phase 1 might have been a better reference standard. It is possible that the increased sensitivity of Quantra for BIRADS 1 and 2 in Phase 2 may be due to the small sample size compared with Phase 1 and the laboratory effect.
Eng 2014[9] and Busana 2016[53]	Cases: women with newly diagnosed breast cancer (mean (SD) age: 67.5 (12.7) years; not eligible as diagnostic population); controls: women who attended routine screening and were found to be breast cancer free (mean (SD) age: 59.5 (6.6) years) (n=1969)	BI-RADS 4th edition; Cumulus v3; ImageJ-based method; Volpara v1.0; Quantra v1.3; single energy x-ray absorptiometry (SXA) method, v6.5	Inter- and intra-method and left-right comparisons among controls. Within-observer reliability of Cumulus. Between-observer reliability of Cumulus. LIBRA	2; UK	7/2/2/0	No	Yes	Yes	The study population was predominantly postmenopausal, thus, limiting the generalizability of the findings to premenopausal women. Response rates were low for healthy controls (51%). Processed images were missing for 15 % of the control participants due to a logistical error.
Eom 2017[45]	Healthy women (n=1000)	BIRADS 5th edition, Volpara version 1.5.12	Intra- and inter-reader agreement for BIRADS; concordance between Volpara and BIRADS	1; Republic of Korea	5/0/4/1	100% Asian	Unclear	Yes	All mammographic examinations performed in a single unit, with only one kind of automated quantitative measurement. Few readers all trained at the same institution. The automated volumetric

									measurement was used as a reference standard. The 5th edition of BI-RADS no longer indicates percentage of dense tissue and emphasises changes in mammography sensitivity. No other gold standard.
Garrido-Estepa 2010[46]	Women aged ≥4 years who attended screening in Barcelona, Burgos, Corunna (Coruña), Palma de Mallorca, Pamplona, Valencia and Zaragoza (n=1532)	BI-RADS 4th edition	Intra-observer reliability	3; Spain	4/1/4/2	Unclear	No	Yes	1 reader only.
Gweon 2013[42]	Full-field digital mammography (FFDM) examinations (n= 778)	BIRADS 4th edition; Volpara version 1.5.1	Inter-rater reliability for BIRADS. Concordance between BIRADS and Volpara	1; South Korea	3/1/6/1	Unclear	Yes	Yes	No reference standard to evaluate breast density. Three radiologists in a single institution assigned BI-RADS density. It would be best to perform a larger study with more patients and radiologists from a variety of practice settings to validate the findings.
Harvey 2013[49]	Women aged ≥ 40 years who underwent ≥2 digital screening mammography examinations <36 months apart; mean (SD) age 57.7 +/- 11.4 (range 40-89 or older) years (n=87066)	BIRADS 3rd edition (prior to 2003) or 4th edition (released in 2003)	BIRADS test-retest agreement	5; USA	4/0/4/3	Yes	Yes	Yes	Included density interpretations determined on both 3rd and 4th editions of BIRADS lexicon

Holland 2016[40]	Women aged 50-75 with consecutive exam pairs; mean (SD) age 58.8 ± 6.7 years (n=500)	Volpara v 1.5.0 and BIRADS 4th edition	Inter-exam agreement was calculated with Cohen's weighted kappa. Intraclass correlation coefficients (ICCs) were calculated to examine the interexam agreement of the four classes categorisation.	Not stated but multiple; The Netherlands	6/1/3/1	Yes	Yes	No	The readers had a minimum of only 1 week between readings (although 30 months between prior and current mammograms). Variability may increase with interval, decreasing agreement over time. In practice agreement might be lower because the screening interval is much longer.
Irshad 2016[12]	Consecutive women with digital mammograms from screening mammography database; mean age 47 (range 36-82) years (n=104)	BIRADS 4th edition and BIRADS 5th edition	Each radiologist evaluated breast density of 104 mammograms four times: twice using the 4th edition BI-RADS criteria and twice using the 5th edition. Intra-reader and interreader agreements for 4th and 5th edition criteria.	1; USA	6/0/4/1	Unclear	Yes	Yes	Readers focused all their attention on breast density, making density the most important finding on the mammograms, which is not the case in real practice in which density is usually a secondary focus of attention.
Irshad 2017[51]	Digital screening mammograms read by the 5 readers at the authors' institution who had read mammograms under 4th (n= 19066) or 5th (n= 16907) edition BIRADS guidelines	BIRADS 4th edition and BIRADS 5th edition	Intraclass correlation coefficient (ICC) within each dataset.	1; USA	3/1/6/1	Yes	Yes	Yes	Single institution; practice patterns of the readers might have been more similar to one another than those seen across various institutions and practices
Jeffers 2017[14]	Cases: women who had screening mammogram and subsequently	Cumulus 6 (version 4.0); Volpara (version not	Correlation between methods	1; USA	2/1/6/2	Unclear	Yes	Unclear	The available sample size limited the ability to detect subtle differences in

	diagnosed with breast cancer; pre-diagnostic mammogram ≥1 year before diagnosis; image of the noncancerous contralateral breast (n=125; 58.4% >50 years). Controls: women without a history of breast cancer who had screening mammogram; breast cancer–free status confirmed with at least 10 years of follow-up for women aged ≥50 years or ≥3 screening mammograms negative for cancer (BI-RADS 1 or 2) for women < 50 years (n=274; 58.8% >50 years).	stated) and BI-RADS (version not stated)							discrimination among the density assessment methods. BI-RADS density assessment by a single reader. Cumulus assessments by a single reader. Using Cumulus requires the reader to undergo specialised training and attain high levels of intrareader reproducibility with test images before reading study images; this and the time required to perform Cumulus measurements made it impractical to have more than one Cumulus reader for this study; having multiple readers could have strengthened the results.
Kang 2016[43]	Craniocaudal (CC) mammograms of subjects who were involved in a breast cancer screening program and found to have normal breasts; mean 50.2 years; range, 28–79 years (n=100)	Cumulus (version 4.0)	Intra- and inter-reader reliability with Cumulus	1; South Korea	4/3/4/0	No	Yes	Yes	The authors chose readers with sufficient experience in mammographic reading and breast density estimation, the small number of readers limits generalisability of findings. They used only CC mammograms. Studies have shown better associations between percent density and breast cancer on CC images than

									on MLO images. Images from one model of equipment. Because each type of mammographic system has different imaging characteristics and post-processing options, results cannot be directly applied to mammograms obtained with other types of equipment.
Kerlikowske 2017[52]	Digital screening examinations of women with incident invasive breast cancers and matched control subjects without prior breast cancer. (n=5406)	BIRADS 5th edition, Volpara version 1.5.0	Correlation between BIRADS categories and Volpara continuous dense breast volume, divided into quartiles	Not stated; USA	5/1/4/1	Yes	Yes	Yes	In studies for interrater and intrarater reliability of the BI-RADS categories, investigators have reported moderate to substantial agreement; misclassification of BI-RADS categories may have influenced results (under- or overestimation of associations). Population predominantly white and Asian; studies should be repeated with Black and Hispanic women to ensure generalisability of results across racial/ethnic groups.
Llobet 2014,[15] Martinez Gomez 2014[54] and Pollan 2013[55]	Mammograms from women participants at two screening centers equipped with full-field digital mammography machines; range 45-69 years (n=655)	BIRADS 3rd edition, DM-Scan, Cumulus	Inter- and intra-rater concordance with DM-Scan and BIRADS. Agreement between visual scale and Cumulus versus DM-Scan, with	2; Spain	5/0/6/0	Yes	Yes	Yes	Brightness correction could introduce a significant error in MD measurement. A hard classification was used, assuming that each pixel can only belong to one of the two possible

			Cumulus/DM-Scan having Concordance Correlation Coefficient (CCC) and Bland-Altman plots.						classes, rather than a soft or probabilistic classification, in which each pixel has a probability of belonging to each class. The authors did not estimate the extra time necessary to add the estimation of breast density to daily routine. DM-Scan and Cumulus were used on processed mammograms that depend on the manufacturers; the authors did not have access to raw (unprocessed) images because Spanish screening centres discard them due to storage constraints. Reliability of DM-Scan and Cumulus not compared.
Lobbes 2012[16]	Women with digital mammograms; mean 51.6 (range 23.9-91.2) years (n=200)	BIRADS 4th edition, QWIN semi-automated thresholding	Inter-reader reliability of BIRADS 4th edition; QWIN ICC left versus right breast	1; The Netherlands	3/0/6/2	Unclear	Unclear	Yes	Included relatively small numbers of dense breasts (BIRADS 3 or 4). A true gold standard for the assessment of breast density is lacking.
Mazor 2016[39]	Patients who had undergone consecutive mammography between January and March 2014 were randomly chosen; age not stated (n=503)	BIRADS 5th edition	Inter-observer agreement between technologists and radiologists. Intra- and inter-observer agreements within the group of radiologists	1; Israel	8/0/2/1	Unclear	Yes	Yes	The reference range for breast density used in this study stemmed from the subjective measurements performed by the radiologists, as methods of objective breast density

			and the inter-observer agreement within the group of technologists.						measurement such as automated breast density measuring algorithms are unavailable in the authors' institution.
Osteras 2016[41] and Osteras 2016[56]	Women with digital mammograms; mean (SD) age 59.3 (5.6) years; range 50-70 years (n=537)	BIRADS 4th edition, Quantra version 2.0 (areometric density, volumetric density, BIRADS-like categories)	Inter-observer variability for each radiologist versus the median BIRADS score (unweighted kappa and with quadratic weights)	1; Norway	4/0/7/0	Unclear	Yes	Yes	The radiologists had a range of experience from 1-34 years, but more- and less-experienced readers equally influence the median score. Radiologists did not use BIRADS in their daily practice but the three categories used in the Norwegian breast cancer screening program. They trained in the use of BIRADS before the study began; the training could reduce the variation in their assessments. This is a single-centre study, using the BIRADS 4th edition, but in the future the 5th edition will be used.
Raza 2016[50]	Digital bilateral screening mammograms; age not stated (n=200)	BIRADS 4th edition; Volpara version not stated	Inter-rater reliability of radiologists using BIRADS before and after training, compared with a) senior breast imagers (leads truth [LT]) and b) Volpara (quantitative truth [QT]).	1; USA	4/1/4/2	No	Yes	Unclear	There is no gold standard for breast density assessment. Today's software is not yet able to account for the complexity of breast tissue, as a trained radiologist can.

Sartor 2016[47]	Digital mammograms with available raw data from the Malmo Breast Tomosynthesis Screening Trial (MBTST), a prospective study comparing MLO DBT alone vs. CC and MLO DM; mean age 58 (range 40-76) years (n=8426).	BIRADS 4th edition and Volpara (version 1.5.11)	Inter-observer variability for examinations with two BIRADS scores. Kappa values for comparison between Volpara density grades (VDG; categorical variable with four groups) and BIRADS scores calculated using separate kappa coefficients for each reader vs. Volpara, then results combined in a meta-analysis, weighting them using the standard error for each kappa, rendering a pooled kappa.	1; Sweden	3/0/8/0	Unclear	Yes	Unclear	Initial trial participation rate was 71.1%; further women did not have both BIRADS and Volpara readings, so overall around 67% participation.
Seo 2013[44]	Healthy women received four-view screening mammograms whose mammograms were considered to be negative (BI-RADS category 1); mean 49.1 (range 35–72) years (n=193)	BIRADS 4th edition and Volpara (version 1.4)	Intra- and inter-observer agreement for the BI-RADS density category; concordance	1; Republic of Korea	5/1/5/0	No	Yes	Yes	There is a lack of reference-standard regarding breast density. Only a small number of radiologists read the BI-RADS breast categories. <30% of eligible women consented.
Singh 2016[38]	Asymptomatic females >35 years of age; mean (SD) 48.8 (7.07), range 36-76 years (n= 476)	BIRADS 5th edition and Volpara (version 1.4.5)	Interobserver agreement using BIRADS; correlation between BIRADS and volumetric breast density	1; India	4/1/3/3	Yes	Yes	Yes	Small single-institution study; examinations were interpreted by only 2 radiologists. No reference standard for breast density. Factors such as BMI were

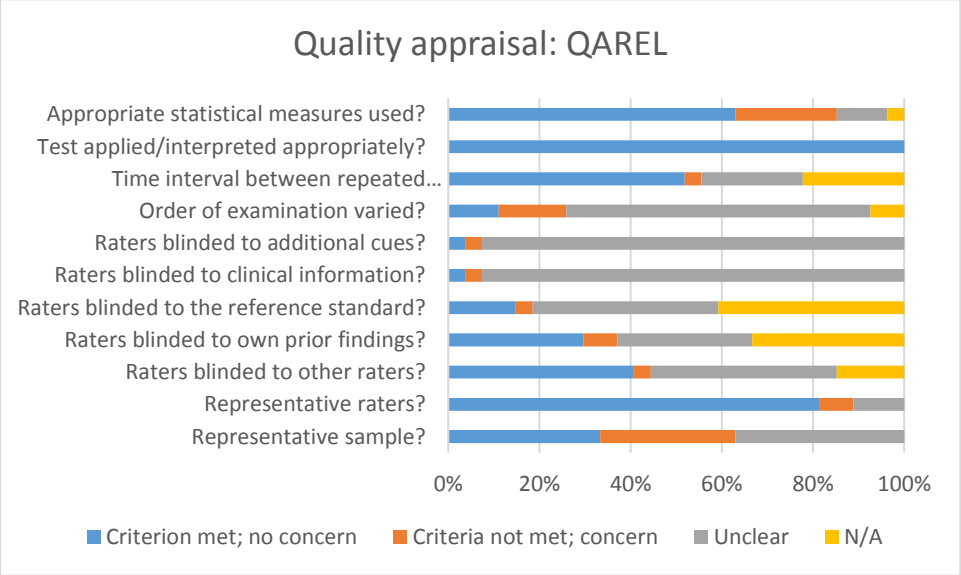
									not investigated. Only one mammography machine was used so results cannot be generalised to all types of machines.
Sprague 2016[22]	Screening mammography; mean (SD) 57.9 (10.8), range 40 to 89 years (n= 145,123)	BI-RADS 4th edition	Inter-rater variation between radiologists; test-retest reliability when interpreted by the same or a different radiologist	30; USA	4/1/6/0	Yes	Yes	Yes	Study limited to assessments by radiologists practicing in the clinical networks of the 3 PROSPR breast cancer screening research centers. Although these included a large number of academic and community practice breast imaging facilities in 4 states, the degree of variation in breast density assessment may differ in other clinical settings around the country, and at radiology practices serving a different demographic mix of patients. Quantitative density measures were not available for comparison with the radiologist's subjective assessment. Results likely reflect not only variation in radiologist interpretation of images but also the variation in the mammography machines and software used to produce digital

									mammographic images that is routinely present across and within facilities over time in clinical practice. Over 15% of women were excluded.
van der Waal 2015[13]	Screening mammograms; median age 59 (IQR: 54–64) years (n=992)	BI-RADS 5th edition; Quantra (version 1.3); Volpara (version 1.5.11)	Intra- and inter-rater reliability of the BI-RADS density scores; overall proportions of agreement (absolute agreement); intraclass correlation coefficients (ICC) between volumetric breast density estimates and BI-RADS classification	1; The Netherlands	6/0/5/0	Yes	Yes	Unclear	The authors did not have information on breast cancer risk, which would ultimately be needed to validate both breast density measures and potentially implement them in a breast cancer screening setting if they are to be used for risk stratification.

Methodological quality of included studies

We found no multi-centre studies that included a representative samples of women and raters, with investigations repeated within the 2-year time-frame. Figure 6 shows the methodological quality appraisal of the included studies. All applied the test criteria appropriately, and most had representative raters, an appropriate time interval between tests and appropriate statistical tests. Blinding was often unclear, and several studies had concerns over statistical measures and the representativeness of the sample. The mean number of criteria met (out of 11) was 4.33 (39%) with a range of 1[17] to 8.[39] The mean number of criteria not met (domains of concern) was 0.96 (0.9%) with a range of 0[12,13,15,16,37,39,41,45,47,49] to 7.[17] The domain which was the most frequent cause of concern (8/27 studies; 30%) was the representativeness of the sample, due to including only women with negative/benign screening results, or only those who went on to have cancer, or over-sampling women with dense breasts. Other domains of concern were statistical measures (identified in 6/27 studies; 22%) and varying the order of examinations (identified in 4/27 studies; 15%). Another issue was unclear or incomplete reporting that prevented an assessment of methodological quality, especially for the blinding domains, varying the order of examinations and the representativeness of the sample. The mean number of domains which were unclear was 4.44 (40%) with a range of 0[17] to 8.[47] The mean number of domains which were not applicable was 1.22 (11%) with a range of 0[9,13,15,19,22,41,43,44,47] to 3.[17,36,38,48,49]

Figure 6. Quality appraisal of included studies for question 1 according to QAREL criteria



Beyond the methodological quality of studies, there are concerns about the applicability to the UK screening population due to the wide age ranges of included women[12,16,39,43] and the different ethnic groups of the included women.[42-45]

3.1.3 Analysis of the evidence

Outcomes reported included intra- and inter-observer reliability of density measurement methods and the concordance between methods, measured using the kappa statistic. While a kappa of 1 represents a perfect agreement, kappa values of 0 or below represent agreements that occur by chance, or that are poor. Kappa values of 0.01–0.20 represent slight agreement, values of 0.21–0.40 represent fair agreement, those between 0.41–0.60 represent moderate agreement, values between 0.61–0.80 represent substantial agreement and values between 0.81–0.99 represent almost perfect agreement.[34] The intra-class correlation coefficient (ICC) is equivalent to the weighted kappa. ICC of less than 0.40 represents poor agreement, 0.40-0.59 represents fair agreement, 0.60-0.74 represents good agreement, and 0.75-1.00 represents excellent agreement.[35]

Six studies assessed reliability and concordance using inappropriate statistical tests: Pearson's correlation coefficients, Spearman's rank coefficients or t-tests.[9,17,38,42,44,52] These are not appropriate measures of reliability or agreement as they either assess linear relationships without detect systematic error (Pearson's, Spearman's) or detect systematic differences but are not sensitive to random difference from the mean (t-test).[57] Therefore, these analyses have been excluded from our results. Appropriate kappa statistics were calculated where possible, if not already presented in the publications.

Analyses examined the kappas for the four density categories (e.g. BIRADS I, II, III, IV) or collapsed into two categories (i.e. dense vs. non-dense).

Test-retest reliability

Visual methods

Overall, test-retest reliability for the four BIRADS categories was moderate to almost-perfect for visual assessment measures of mammographic breast density. The percentage agreement between raters on BIRADS versions 3, 4, and 5 was reported in nine studies, and kappa ranged from 0.54 to 0.95. For BIRADS 3rd or 4th edition, one study showed moderate test-retest reliability ($\kappa = 0.54$; not stated to be weighted) on the four-category scale.[49] We calculated the weighted linear kappa as 0.638 (95% CI 0.634, 0.642) indicating moderate agreement. For BI-RADS 4th edition, one study[19] reported weighted kappas (weighting not stated) for three radiologists (0.86, 0.87 and 0.88) indicating almost perfect agreement on the four-category scale and weighted kappas on the two-category scale of 0.88, 0.90 and 0.91. One study[46] reported a quadratic weighted kappa of 0.90 for one radiologist indicating almost perfect agreement on the four-category scale and 0.82 on the two-category scale. One study[40] reported weighted kappa values (weighting not stated) for three radiologists (0.76, 0.77 and 0.79) indicating substantial agreement and a PhD student with a medical degree and two years of experience with breast imaging (0.82) indicating almost perfect agreement on the four-category scale, and 0.68–0.77 on the two-category scale. One study[12] reported individual intrareader agreements (quadratic weighted kappa) in five radiologists ranged from 0.78 to 0.92;

four readers scored >0.8 indicating almost perfect agreement and one 0.78 indicating substantial agreement on the four-category scale. One study[22] involved 83 radiologists and we calculated the linear weighted kappa of 0.760 (95% CI 0.7507, 0.7695) indicating substantial agreement and quadratic weighted kappa of 0.8338 (95% CI 0.8172, 0.8504) indicating almost perfect agreement for the two-category scale.

For the most recent 5th edition of BI-RADS, test-retest reliability was moderate to almost-perfect in 3 studies (κ = 0.74–0.95).[12,13,45] In one study, two breast-imaging experts (kappas 0.84, 0.87), two general radiologists (kappas 0.86, 0.95), and one student (kappa 0.86) had almost perfect agreement, while one student's agreement was substantial (0.74) on the four-category scale.[45] Intra-reader agreements on the two-category scale were almost perfect or substantial (k=0.76–0.95) among the breast-imaging experts (0.85, 0.88), general radiologists (0.88, 0.95), and students (0.76, 0.90). One study[12] reported individual intrareader agreements (quadratic weighted kappa) in five radiologists ranged from individual intrareader agreements in five readers ranged from 0.74 to 0.99; four readers scored >0.8 indicating almost perfect agreement and one 0.74 indicating substantial agreement on the four-category scale. One study[13] reported quadratic weighted kappas for three radiologists (0.82, 0.85 and 0.87) on the four-category scale.

Semi-automated methods

The semi-automated DM-Scan was assessed in one study and test-retest reliability was almost-perfect in three radiologists (ICC 0.900, 0.935 and 0.938; mean of the three readers: 0.924) on the four-category scale.[15]

Fully-automated methods

One study assessed the fully-automated Volpara using serial mammograms over time and test-retest reliability was almost-perfect (weighted κ = 0.85; weighting not stated) on the four-category scale and 0.80 on the two-category scale.[40]

Inter-rater reliability

Visual methods

Overall, inter-rater reliability was fair to almost perfect for visual methods. The agreement between raters on visual percent density was assessed in one study comparing four readers (ICC [equivalent to a quadratically weighted kappa] = 0.884).[37] The BI-RADS 4th edition was assessed in ten studies. One study[19] reported a weighted kappa (weighting not stated) between pairs of radiologists of 0.66, 0.73 and 0.75 on the four-category scale and 0.77, 0.83 and 0.89 on the two-grade scale. One study[42] reported the overall weighted kappa (weighting not stated) of the three radiologists' estimates of BI-RADS density categories showed moderate agreement (κ = 0.48). One study[40] reported weighted kappa values (weighting not stated) between 0.78 and 0.83 for the four-category scale and between 0.73 and 0.78 on the two-category scale between three radiologists and a PhD student with a medical degree and two years of experience with breast imaging. One study[12] reported an overall interreader agreement (quadratic weighted kappa) of 0.65, with quadratic weighted kappa between

pairs of radiologists of 0.67, 0.71, 0.74, 0.75, 0.77, 0.80, 0.82, 0.84, 0.86 and 0.87. One study[51] reported an ICC between five radiologists of 0.940. One study[15] reported an average quadratic weighted kappa of 0.823 between three radiologists. One study[41] reported that four of the five radiologists had almost perfect agreement with the median score using quadratic weights (0.849, 0.875, 0.879, 0.934) and the fifth had substantial agreement (0.763). One study[47] reported a linear weighted kappa of 0.77 between five radiologists. One study compared a breast radiologist with 18 years' experience versus a senior resident in radiology with 2 years' experience (overall linear weighted κ = 0.521 [reported by study authors] indicating moderate agreement; quadratic weighted κ = 0.65, 95% CI 0.53, 0.77 [calculated by us] indicating substantial agreement).[16] Results from the largest multi-centre real-world setting study[22] showed that:

- Among women with consecutive mammograms interpreted by different radiologists (n = 34 271 women), at a median interval of 1.1 years (IQR 1.0 to 1.3 years), 27.0% of women with dense breasts at the first examination were classified as nondense breasts at the second examination, and 11.4% of women with nondense breasts at the first examination were classified as dense breasts at the second examination. Differences between radiologists persisted after adjustment for age, race and BMI.

- The median percentage of mammograms rated as showing dense breasts was 38.7% (IQR 28.9% to 50.9%; range 6.3% to 84.5%). A quarter of radiologists rated <28.9% of their patients' mammograms as showing dense breasts, whereas the highest 25% of radiologists rated at least 50.9% of their patients' mammograms as showing dense breasts.

- There was substantial variation across radiologists in the percentage of mammograms rated as showing dense breasts within nearly all age and BMI categories.

Seven studies assessed the BIRADS 5th edition. One study[48] reported weighted kappas (weighting not stated) between pairs of readers of 0.56, 0.59; and 0.68 on the four-category scale and 0.67, 0.67 and 0.82 on the two-category scale. One study[36] reported unweighted kappas between pairs of readers of 0.38, 0.58 and 0.68 on the four-category scale and 0.70, 0.81 and 0.85 on the two-category scale. One study[12] reported an overall interreader agreement (quadratic weighted kappa) of 0.57, with quadratic weighted kappa between pairs of radiologists of 0.61, 0.72, 0.74, 0.75, 0.76, 0.77, 0.79, 0.85, 0.85 and 0.90. One study[38] reported almost perfect agreement (weighted κ = 0.895; weighting not stated) for two blinded radiologists. One study van der Waal 2015[13] reported a quadratic weighted kappa of the inter-rater comparisons of three radiologists ranged from 0.80 to 0.84 for the four-category scale and 0.89 to 0.90 for the two-category scale. One study compared the agreement between breast-imaging experts with more than five years of experience in reading mammograms versus two general radiologists with fewer years of experience in reading mammograms (weighted κ = 0.67 on the four-category scale; 0.78 on the two-category scale; weighting not stated), even though for inter-reader analysis, the reader with better intra-reader agreement was chosen from each group.[45] One study[39] compared ten mammography technologists (weighted kappa 0.62 within this group on both the four-category scale and the two-category scale) and seven breast radiologists (weighted kappa 0.69 within this group on the four-category scale and

0.77 on the two-category scale). Only a fair level of agreement was noted between the technologists and the radiologists (weighted kappa 0.38 between groups on the four-category scale and 0.45 on the two-category scale).[39]

Semi-automated methods

Two studies assessed Cumulus using radiologists, breast surgeons or the reader profession was not stated (κ = 0.83–0.90).[9,43] One study[9] reported the ICC 0.89, 0.90 and 0.83 for raw ("for processing"), processed ("for presentation") and analogue-like images, respectively. One study[43] reported a concordance correlation coefficient (CCC) of 0.86-0.89 between two radiologists board certified in breast imaging and one breast surgeon. One study assessed the semi-automated DM-Scan used by radiologists and reported ICC between pairs of readers of 0.916, 0.922 and 0.928.[15]

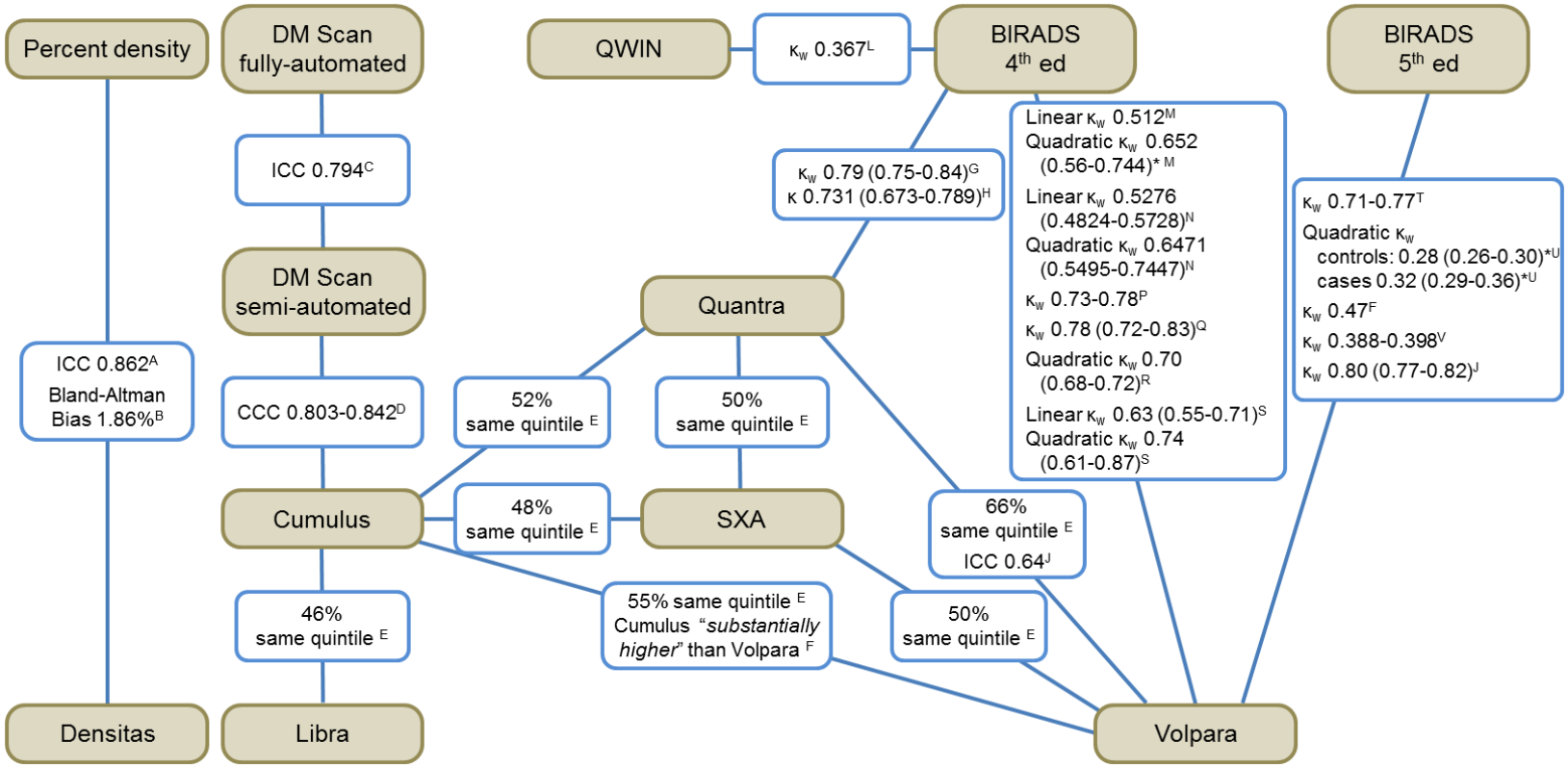
Fully-automated methods

Inter-rater reliability is not applicable for fully-automated measures as they do not require human raters.

Concordance

Concordance between methods was assessed in 17 studies and agreement varied from fair to substantial: three studies reported kappa between 0.21 and 0.40 (fair agreement); one study between 0.41 and 0.60 (moderate agreement); twelve studies between 0.61 and 0.80 (substantial agreement) and two studies between 0.81 and 0.99 (almost perfect agreement) (see Figure 7). One study compared the quintiles of density defined by different methods; the highest concordance between pairs of methods was for Quantra and Volpara, but even for this pair (but fully automated volumetric methods), only 66% of women were assigned to the same quintile.[9]

Figure 7. Diagram of concordance (excluding untrained students)



* Kappa calculated

CCC = Concordance Correlation Coefficient ICC = Intraclass correlation coefficient; κ = Unweighted Kappa; κ_w = Weighted Kappa

A: Abdolell 2013
B: [*limits of agreement -20.38 to +24.1, largest outlier not reported*]
C: Llobet 2014
D: Pollan 2013
E: Eng 2014 & Busana 2016
F: Jeffers 2017
G: Ekpo 2016 [*Quantra…*]
H: Osteras 2016 [*Classification...*]
J: van der Waal 2015
L: Lobbes 2012. [*Experienced reader*]
M: Destounis 2017
N: Gweon 2013
P: Holland 2016
Q: Raza 2016
R: Sartor 2016
S: Seo 2013
T: Eom 2017
U: Kerlikowske 2017
V: Singh 2016

3.1.4 Discussion

Study evidence

The likelihood of a woman being told she has dense breasts varies substantially within and between readers for visual methods (see Table 4). Semi-automated and automated methods are more consistently reliable than visual methods. However, although semi-automated methods have been shown to have high between- and within-reader reliability in research settings, in which efforts are made to train the readers and ensure standardisation of procedures, similar high inter-reader reliability values may not be achieved in clinical practice.

Table 4. Reliability and validity measures (kappa, ICC) for different types of density assessment methods (NB kappa values between 0.81–0.99 represent almost perfect agreement).

	Visual	Semi-automated	Automated
Test-retest	0.54-0.95	0.92	0.85
Inter-rater	0.38-0.96	0.83-0.92	

Note that a difficulty with immediate test-retest assessment in mammography is that because of the radiation dose associated with mammography, a good reason is required to repeat the mammograms, either in the same compression, or in a different one; test-retest over time is a proxy measure.

Concordance between methods also varied (see Table 5) and is not generally high, as methods define density in different ways and there is no gold standard applicable to all breast density measurements. Even automated methods such as Volpara and Quantra clearly differed from each other, i.e. methods are not interchangeable.

Table 5. Concordance between methods

	Semi-automated	Automated
Visual	-	"Significantly different" to almost perfect agreement; 0.28-0.86
Semi-automated	Almost perfect agreement: 0.80-0.84	"Substantially different" to substantial agreement; 0.79; 46-52% assigned to the same quintiles
Automated	-	Substantial agreement: 0.64; 50-66% assigned to the same quintiles

Study quality

High quality studies would have low risk of bias and should also be generalisable to our population in terms of the women (a large number of representative women from a general screening population) and the readers (a large number of readers within a multi-centre study of general screening, rather than single centre studies or readers specially trained for a research study). None of the studies scored above 8/11 for domains of the quality assessment tool that were met (no concern), and even those studies with most of the domains met had domains not met or unclear.

Study applicability

Although most studies included a sample that was representative of a UK screening population, there are concerns about the applicability of some of the studies to the UK screening population due to the wide age ranges of included women[12,16,39,43] and the different ethnic groups studied, for example in the studies conducted in the Republic of Korea.[42-45]

Consistency

The studies consistently showed that repeatability of density measurements was higher for the same reader than for different readers using the same measurement method, and lower for concordance studies comparing different measurement methods.

3.1.5 Summary

This question addressed NSC criterion 4: There should be a simple, safe, precise and validated screening test. **Not met.**

It is difficult to validate the density methods when there is no gold standard against which to compare breast density measurements, concordance between methods is variable. It is clear that even automated methods are not interchangeable.

3.2 Key questions 2a and 2b

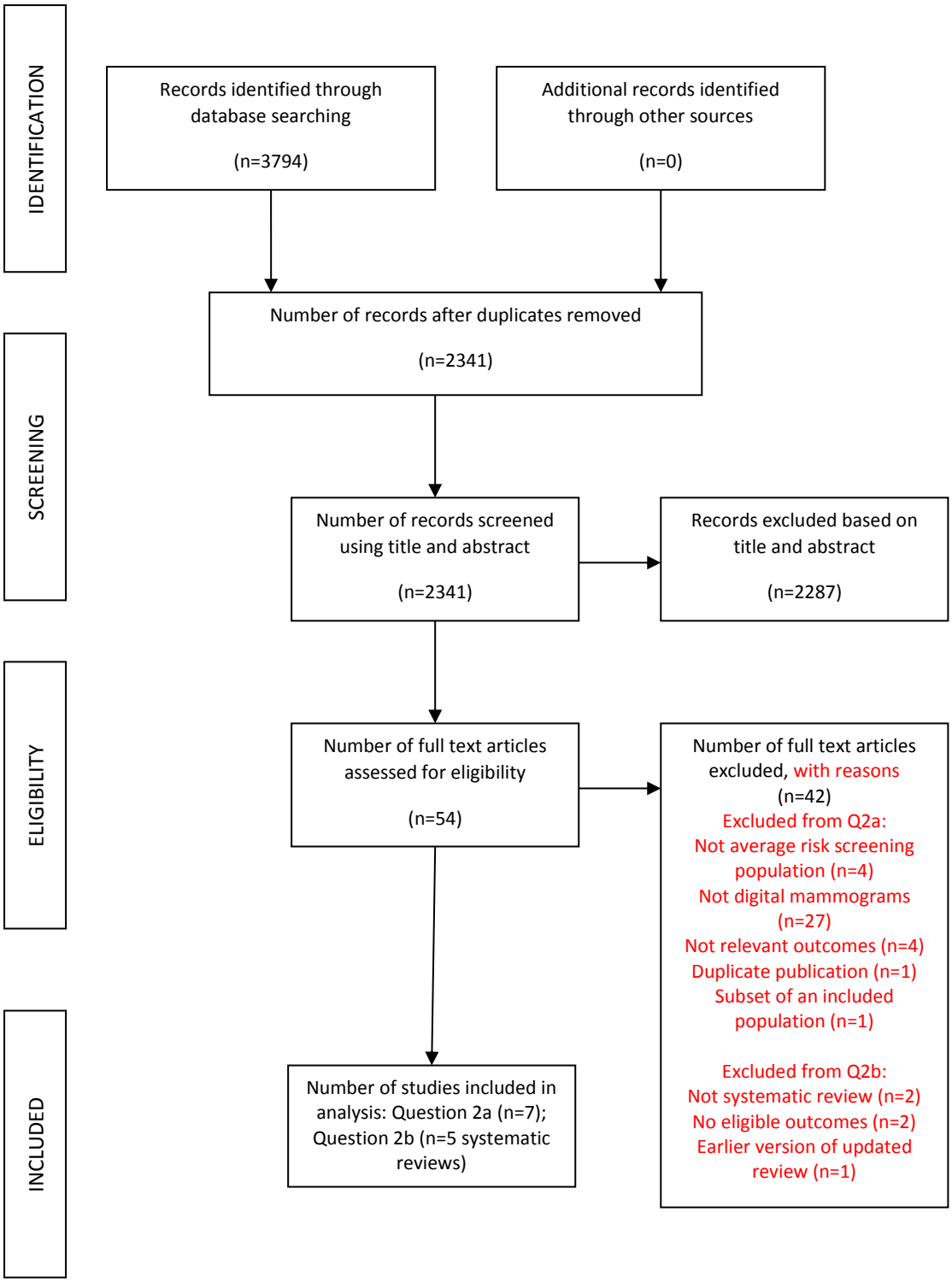
2a: Is mammographic breast density a risk factor for cancers being missed during screening (masking on mammograms/false negatives/interval cancers)?
2b: Is mammographic breast density a risk factor for developing breast cancer?

These relate to NSC criterion 1:
"There should be robust evidence about the association between the risk or disease marker and serious or treatable disease."

3.2.1 Description of the evidence

The searches identified 3794 studies through electronic databases; 261 records were examined at title and abstract stage, of which 54 were examined as full texts. Seven studies were subsequently included for question 2a, and five studies for question 2b. Details of the excluded papers are provided in Appendix 3. The numbers of papers at each stage of the search are shown in the PRISMA flow chart below (see Figure 8).

Figure 8. PRISMA flowchart for question 2



3.2.2 Question 2a

Characteristics of included studies

Seven studies were included (see Table 6 and Appendix 6). Sample sizes ranged from 60[58] to 405,191[59]. The studies were conducted in Australia,[58] Belgium,[60] the Netherlands[7,61] and the USA.[18,59,62] Visual density methods (BIRADS) were used in six studies;[18,58-62] an automated method (Volpara) was used in three studies.[7,18,61]

Table 6. Characteristics of included studies

Study	Population (n)	Interventions/ Comparator	Outcome	No. centres; country	Limitations
Destounis 2017[18]	Women aged >40 years (mean 62.1; SD 11) with histopathologically confirmed breast cancer (n=614)	Mammographic density using BIRADS 4th edition or Volpara	Comparison between screen-detected and interval cancers	1; USA	Retrospective study; BMI not available and so not included in multivariate analysis. Interval cancers not differentiated between true interval, missed or mammographically occult (i.e. masked by dense tissue). Unable to analyse the relation between masking risk and location and distribution of density within the breast. Large proportion of people missing from analysis. Around 13.6% aged <50 years and 23.6% >70 years. Around 8.5% <47 years and 16.1% >73 years.
Holland 2017[61]	Cases: Women with interval cancers within 12 months after the examination. The last available screening examination before cancer diagnosis is used in this study. Mean age 57.7 years. Controls: For each patient with an interval cancer, 10 participants were chosen as controls. The control participants needed to have had a mammographic examination in the same month in which the last screening examination of the	Percent dense volume using Volpara or percent density using BIRADS 5th edition	To measure to what extent the methods can identify women at high masking risk, the mammograms were divided in a high and low masking risk group by thresholding the risk measure. Then, the sensitivity of the masking measures was computed as the number of interval cancers in the high-risk group divided by the total number of interval cancers. The false positive rate is calculated as the percentage of normal controls	1; The Netherlands	Given that the exact cancer location was unknown and that diagnostic mammograms were not available, it was not possible to review the interval cancers and to confirm that masking is the cause for a cancer diagnosis outside the screening program. CC images not available for all exams. BIRADS density assessments of only one radiologist. Many studies found inter- and intra-reader variability in breast density assessment using BI-RADS.

	interval cancer patient was performed. To be eligible as control, the women should not have been recalled on the basis of this mammographic examination and they should not have been diagnosed with breast cancer within 2 years after this examination. Controls without a density map, due to failure of the computation, were replaced. (n=111 cases + 1110 controls). Mean age 59.2 years.		selected as at high masking risk at the same threshold. In the context of risk stratification for supplemental screening, the proportion of controls selected as at high masking risk can be seen as supplemental screening rate and the proportion of interval cancers gives an estimate about the cancers that might be detectable with additional imaging at that supplemental screening rate.		Therefore, to make a definitive comparison between the automated methods and radiologists assessments, an extensive reader study should be conducted with multiple readers.
Kerlikowske 2015[62]	Women aged 40-74 years who did not have a history of breast cancer or breast implants and had complete information on demographic and breast health history information (n=365,426)	Mammographic density using BIRADS	Interval cancer rate and false positive rate by breast density	Not stated; USA	The cut-points used for defining low performance were developed for identifying minimally acceptable performance levels for screening mammography interpretation for invasive and DCIS outcomes combined; the authors state that they do not know if these performance cut-points are related to long-term outcomes such as breast cancer mortality. For some subgroups with an average interval cancer rate <1/1,000 mammograms, they cannot rule out a higher interval cancer rate because the upper 95% confidence limit exceeds one. A 24-month interval was not evaluated since women may return early for screening and/or have

					mammograms outside the BCSC. Participation rate not stated. 19.1% aged 40-49 years and 13.4% aged 70-74 years
Nelson 2016[59]	Women aged 40 to 89 years who had routine screening with digital mammography (n=405,191)	Mammographic density using BIRADS 4th edition	Rates of false-positive and false-negative mammography results and recommendations for additional imaging and biopsies from a single screening round	5 registries; USA	The BCSC data reflect opportunistic screening in a fluctuating population of women in the USA whose information was collected by the participating registries. Findings may not be applicable to other populations. Restrictions of registry data with pre-defined data elements and the inherent biases of observational data. Some outcomes, such as the effectiveness and harms of different screening intervals, would be more accurately determined by comparing outcomes between women who were randomly assigned to comparison groups. 16.3% had missing data for breast density. 28.1% aged 40–49 years, 12.4% aged 70–79 years and 4.6% aged 80–89 years.
Rawashdeh 2013[58]	A single-image bank containing 60 digital cases containing 20 positive (biopsy-proven) cases with a single focus of cancer in 16 cases and multicentric cancer in 4 cases (resulting in a total of 24 cancers)	BIRADS 3rd edition	Detectability of lesions by breast density in a reader study	Not stated; Australia	The same radiologist who chose the images was responsible for assessing breast density; <100 images

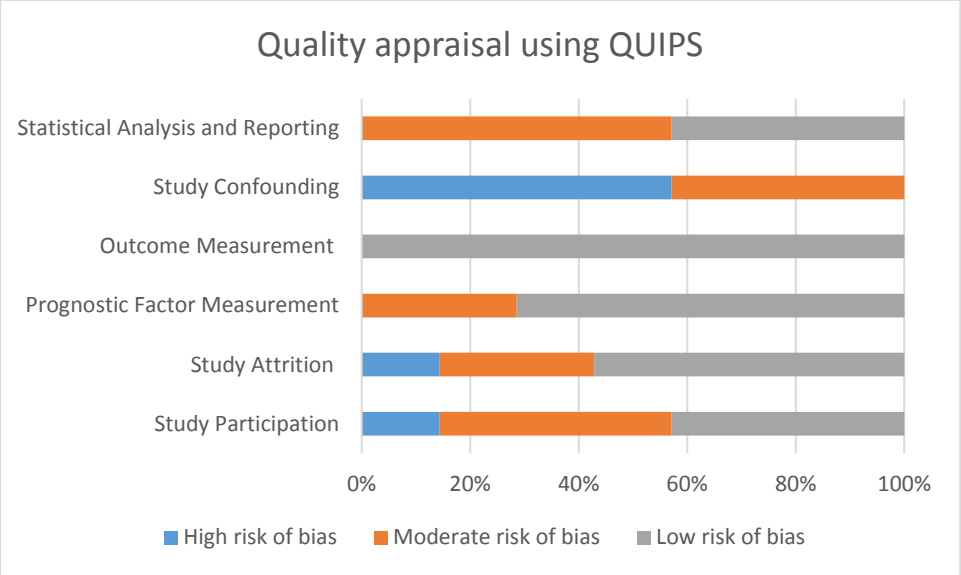
	(n=60). Mean 54 years (range 47 to 78 years)				
Timmermans 2017[60]	Women aged between 50 and 69 years (n=351,532)	BI-RADS 4th edition	Cancer detection rate, interval cancer rate, third readings and correlated false-positives by breast density category	Not stated; Belgium	Subdivision of ICs in true, missed and minimal signs was not performed. A low statistical power hampered reaching statistical significance in differences between modalities for the BI-RADS IV class data.
Wanders 2017[7]	Women aged 50–75 years participating in a biennial screening program (n=111,898 examinations belonging to 53,239 women)	Volpara	Interval cancers by density	1; The Netherlands	The MLO view was the standard view for the subsequent screening rounds and CC views were only taken in addition to MLO during the first screening round or by indication during subsequent rounds. As a result, breast density was determined based on only MLO views for some examinations and on both MLO and CC views for others. Volpara's volumetric percent density measured on CC views tends to be somewhat higher than on MLO views. As CC views are more often performed among women with dense breasts and women with a suspicious region on their MLO view, breast density might be somewhat artificially elevated for these women. Screening sensitivity is presumably higher when both MLO and CC views are available versus MLO views only. Therefore, standardly taking both MLO and CC

					views would lead to higher sensitivity, particularly in women with fatty breasts as they are the ones who most often receive MLO views only. This would lead to larger differences in screening performance across breast density categories.

Methodological quality of included studies

The quality of the included studies is shown in Figure 9. Key quality issues included interval cancers not differentiated between true interval, missed or mammographically occult (i.e. masked by dense tissue);[60] many women missing from the analysis;[18] missing data for breast density;[59] and lack of detail on the included population.[58] Most participating women were aged between 47 and 73 years, although in several studies[18,59,62] over 10% of women fell outside this age range.

Figure 9. Quality appraisal for included studies in question 2a according to QUIPS



Analysis of the evidence

Visual methods

Destounis 2017[18] analysed 614 women aged >40 years (mean 62.1; SD 11) with histopathologically confirmed breast cancer, comparing those with screen-detected and those with interval cancers in 1 centre in the USA. Around 13.6% aged <50 years and 23.6% >70 years. The mammographic sensitivity was reported by BIRADS density and was lower for women with extremely dense breasts: fatty replaced: 82%; scattered fibroglandular: 90%; heterogeneously dense: 84%; extremely dense: 66%; R^2 = 0.463. In univariate analysis, density was associated with the risk of diagnosis of interval cancer versus screen-detected cancer: BIRADS category 3 vs. 1 or 2: OR 1.91 (1.07-3.40), p=0.028; BIRADS category 4 vs. 1 or 2: OR 5.00 (2.43-10.33), p<0.001. In age-adjusted analysis, BIRADS 3 vs. 1 or 2: the OR was 1.60 (0.89-2.89), and for BIRADS 4 vs. 1 or 2, the OR was 3.82 (1.82-8.06), p<0.001.

Holland 2017[61] analysed 111 women with interval cancers diagnosed within 12 months of screening (the last available screening examination before cancer diagnosis was used in this study) versus 1110 control women (who had a mammogram in the same month in which the last screening examination of the case was performed and were not recalled or diagnosed with breast cancer within 2 years after this examination). Percent dense volume using Volpara (see fully-automated section below) or percent density using BIRADS 5th edition were used, in 1 centre in The Netherlands. With BI-RADS, 427/1110 = 38.5% (95% CI 35.7–41.3) of the controls (no cancer) were at increased masking risk,

compared with 70/111 = 63.0% (95% CI 53.5–72.0) of the women developing interval cancers, giving a RR of dense breasts among those with interval cancer of 63/38.5 = 1.64 (calculated by us).

Kerlikowske 2015[62] included 365,426 women aged 40-74 years who did not have a history of breast cancer or breast implants and had complete information on demographic and breast health history information in the USA. The rates of interval cancers increased by density at all ages (see Table 7).

Table 7. Interval cancer rate per 1000 mammograms (95% CI).

	BI-RADS mammographic breast density			
Age (years)	**Almost entirely fat**	**Scattered fibroglandular densities**	**Heterogeneously dense**	**Extremely dense**
40 – 49	0.19 (0.04, 0.56)	0.26 (0.16, 0.40)	0.76 (0.61, 0.93)	0.98 (0.67, 1.37)
50 – 59	0.14 (0.05, 0.34)	0.33 (0.23, 0.45)	0.80 (0.65, 0.98)	1.11 (0.72, 1.64)
60 – 69	0.23 (0.10, 0.45)	0.49 (0.37, 0.65)	0.96 (0.75, 1.22)	1.13 (0.54, 2.09)
70 – 74	0.35 (0.10, 0.90)	0.55 (0.33, 0.86)	1.15 (0.73, 1.72)	3.45 (1.27, 7.50)

Nelson 2016[59] studied 405,191 women aged 40 to 89 years who had routine screening with digital mammography in 5 registries in the USA, using the BIRADS 4th edition. Women with less dense breasts had lower rates of false-negative mammography results than those with more dense breasts (See Table 8).

Table 8. Rates of false-negative digital mammography per 1,000 women screened per round and 95% CI)

	40-49 years	50-59 years	60-69 years	70-79 years	80-89 years
Fat	0.2 (0.0, 0.9)	0.3 (0.1, 0.7)	0.6 (0.2, 1.5)	0.3 (0.1, 1.1)	0.4 (0.1, 3.1)
Scattered	0.5 (0.3, 0.7)	0.7 (0.5, 0.9)	0.8 (0.6, 1.2)	1.2 (0.7, 1.9)	1.0 (0.6, 1.7)
Heterogeneous	1.3 (1.0, 1.7)	1.4 (1.0, 2.0)	1.7 (1.3, 2.3)	2.3 (1.6, 3.4)	1.1 (0.5, 2.4)
Extreme	1.7 (1.2, 2.5)	1.6 (0.9, 2.8)	1.2 (0.6, 2.7)	5.6 (2.4, 12.9)	6.9 (2.5, 18.5)
p value for trend across density groups	<0.001	<0.001	0.02	0.002	0.17

Rawashdeh 2013[58] studied the detectability of lesions by mammographic breast density in a reader study in Australia. The series contained 60 digital cases containing 20 positive (biopsy-proven) cases; women were a mean of 54 years old (range 47 to 78 years). The same radiologist who chose the images was responsible for assessing mammographic breast density using BIRADS 3rd edition. There was a negative correlation between lesion detection on mammography and breast density (r = -0.64, p = 0.007), suggesting that cancers were harder to see on mammograms from women with dense breasts.

Timmermans 2017[60] assessed 351,532 women aged between 50 and 69 years using the BI-RADS 4th edition in Belgium. They found a systematic increase of interval cancer rate with breast-density class: BIRADS I: 1.11 per 1000; BIRADS II: 2.02 per 1000; BIRADS III: 3.80 per 1000; and BIRADS IV: 5.36 per 1000. The percentage of cancers detected in the screening programme over the total number of cancers registered (screen-detected plus interval cancers, reflecting the sensitivity of the screening

programme) decreased from 84% for BIRADS I, to 74% for BIRADS II, to 60% for BIRADS III, to 46% for class IV.

Semi-automated methods

No eligible studies were found.

Automated methods

Destounis 2017[18] reported mammographic sensitivity by Volpara automated density grade: Grade 1: 95%; Grade 2: 89%; Grade 3: 83%; Grade 4: 65%; R^2 = 0.914. Destounis 2017[18] also reported that in univariate analysis, density was associated with the risk of diagnosis of interval cancer versus screen-detected cancer:

- Automated density grade 3 vs. 1 or 2: OR 1.94 (95% CI 1.10-3.43, p=0.021).
- Automated density grade 4 vs. 1 or 2: OR 5.60 (95% CI 2.99-10.47, p<0.001).
- Volumetric breast density quartile 2 vs. quartile 1: OR 1.73 (95% CI 0.72-4.13, not significant).
- Volumetric breast density quartile 3 vs. quartile 1: OR 2.08 (95% CI 0.90-4.83, not significant).
- Volumetric breast density quartile 4 vs. quartile 1: OR 5.58 (95% CI 2.61-11.93, p<0.001).

After adjustment for age, the odds ratios were:

- Automated density grade 3 vs. 1 or 2: OR 1.64 (95% CI 0.92-2.94, not significant).
- Automated density grade 4 vs. 1 or 2: OR 4.14 (95% CI 2.13-8.03, p<0.001).
- Volumetric breast density quartile 2 vs. quartile 1: OR 1.67 (95% CI 0.70-4.01, not significant).
- Volumetric breast density quartile 3 vs. quartile 1: OR 1.85 (95% CI 0.79-4.33, not significant).
- Volumetric breast density quartile 4 vs. quartile 1: OR 4.17 (95% CI 1.89-9.21, p<0.001).

Holland 2017[61] reported that if the thresholds of Volpara percent dense volume were set so that 38.5% of controls were classified as having dense breasts, then 66.1% (CI 55.8–76.2) of the women with an interval cancer had dense breasts.

Wanders 2017[7] studied women aged 50–75 years participating in a biennial screening program (analysed n=111,898 examinations belonging to 53,239 women) in 1 centre in The Netherlands. There was a reduced mammographic sensitivity (%) by breast density (Volpara density grade [VDG]): VDG 1: 85.7% (78.1; 91.0); VDG 2: 77.6% (73.2; 81.5); VDG 3: 69.5% (64.1; 74.4); VDG 4: 61.0% (51.2; 70.0); p<0.001. Interval breast cancer rates were higher in higher breast density categories compared to lower density categories with a significant linear trend (p-trend<0.001). Interval cancer rates in the first year after a screening examination were 0.2, 0.8, 1.2, and 2.9% (p-trend<0.001) in VDG categories 1, 2, 3, and 4, respectively. The interval cancer rate per 1000 was: VDG1: 0.7 (0.4; 1.1); VDG 2: 1.9 (1.5; 2.3); VDG 3: 2.9 (2.3; 3.5); VDG 4: 4.4 (3.2; 6.0); p<0.001.

3.2.3 Question 2b

As several systematic reviews were found in the search for question 2b, it was decided to conduct a systematic review of these systematic reviews (as specified in the protocol). The methods used were those advocated in Smith et al (2011): "Methodology in conducting a systematic review of systematic reviews of healthcare interventions".[63]

Characteristics of included studies

The included studies are shown in Table 9; latest search dates of the systematic reviews ranged from January 1, 2008[64] to December 31, 2015.[65] The number of included studies ranged from five[64] to 37.[66] One systematic review[65] included Asian women only, and in one the age range in included studies was 40-84 years; in the other three systematic reviews the population was not stated. Systematic reviews were assessed for the extent to which they matched our scope; all the included reviews appeared to answer an appropriate question and all included density measurement methods specified in our review protocol. They reported unadjusted outcome and/or age-adjusted outcome measures, or did not report adjustment.

Table 9. Characteristics of included studies

	Our scope:	Bae 2016[65]	Huo 2014[66]	Elias 2014[67]	Antoni 2013[68]	Cummings 2009[64] and McCormack 2006[69]
Question	Q2b: **Is mammographic breast density a risk factor for developing breast cancer?**	This meta-analysis investigated the association between breast density in mammography and breast cancer risk in Asian women.	To critically review the current literature on mammographic density (MD) and summarize the current evidence for its association with breast cancer (BC).	Features (including density) related to HER2 overexpression (a marker of cancer aggressiveness)	A systematic review of studies of mammographic density (MD) in relation to risk of subtype-specific breast cancer, by ER, PR, and HER2 status or gene expression profiles.	To review prospective studies about models and sex hormone levels to assess breast cancer risk and use meta-analysis with random effects models to summarize the predictive accuracy of breast density.
Population	Women aged 50-70 attending breast cancer screening from the general population (not specifically chosen high-risk groups) with a population prevalence similar to the UK	Asian women. Seven datasets were of premenopausal women and eight were of postmenopausal women	Not stated	Not stated	Age range in included studies 40-84 years	Not reported
Density measurements	BI-RADS scale scored by a single qualified reader BI-RADS scale scored by a group consensus of readers • Volpara • Quantra • Cumulus • ImageJ	Wolfe classification; percent density (%); DA, density area (cm^2); MDA, mean dense area (cm^2); TBA, total breast area (cm^2); VDG, volumetric density grade (%); ADA, absolute dense area (cm^2).	BIRADS, Cumulus, Boyd semi-quantitative scale, computer-assisted method (CAM), Tabar, DM-Scan, automated volumetric breast	BI-RADS	BIRADS, percent density, visual (fatty, mixed/dense), Wolfe or Cumulus in different included studies	One study assessed breast density by use of BI-RADS ratings and four measured percent density, in addition to the studies included in McCormack 2006[69]

	• Single energy x-ray absorptiometry (SXA) • DM-Density M-Vu Breast Density • Absolute fat volume • Absolute fibroglandular volume • Density calculated on a single mammogram view (e.g. MLO) • Density calculated from 2 views (e.g. MLO plus CC)		density, automated measure, percent density, semi-automated technique: threshold technique (TT), fully automated method (FAM), semi-automated method (SAM), standard mammogram form (SMF)			
Outcomes	Head to head studies (2 or more types of density measurement): Positive and negative concordance between pairs of tests; comparison of characteristics of discordant cases: in particular comparison of risk of breast cancer and measures of missing cancers at screening such as interval cancers. Single or head to head studies (1 or more types of test): Proportion of women who have an interval cancer after screening by density for each test; proportion of women who have breast cancer by density for each test (includes reporting of absolute risk which is of particular interest in low density	Effect size based on adjusted odds ratios (adjustment factors not stated)	Mammographic density as a risk factor for breast cancer; association of mammographic density with breast cancer subtypes and tumour characteristics.	Odds ratio of HER overexpression by density categories	Relative risk estimates and their 95% CIs of subtype-specific breast cancer were estimated by individual studies as odds ratios in case–control and case-only studies and as hazard/rate ratios in cohort studies. The most fully adjusted RRs reported were included. Controlling for age was included in eligibility criteria. In case-only studies, we extracted estimates of the ratios of relative risks (RRR) of ER+ versus ER- breast cancer	Relative risk of breast cancer; all adjusted for age; some studies adjusted for additional factors which were not stated except to say that studies that further adjust for body mass index or weight observed somewhat stronger associations

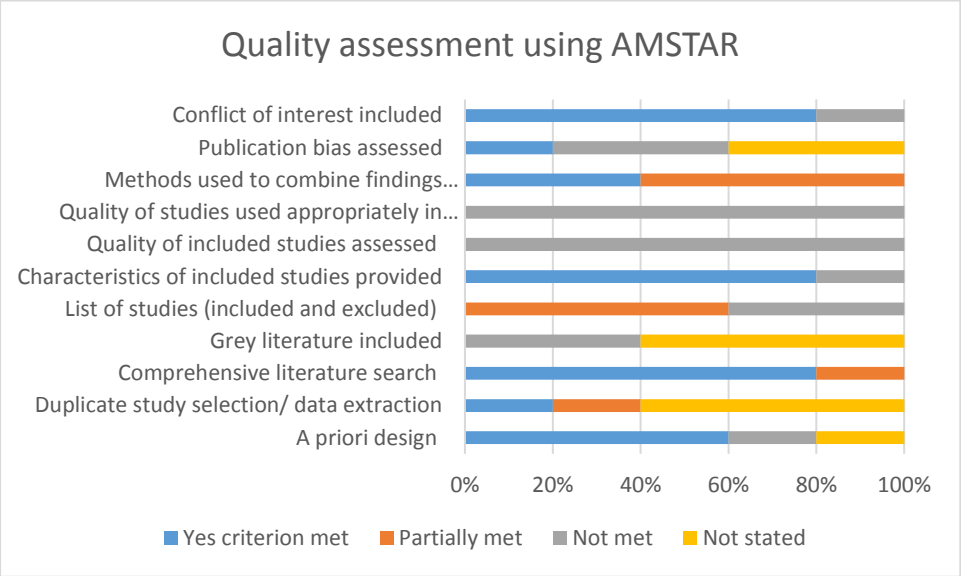
	groups); distribution of cancer type by risk group for each test; Odds or risk ratios from unadjusted univariable models of density as a predictor of risk; odds or risk ratios from age-adjusted multivariate models of density as a predictor of risk				associated with MD categories; if ER+ subtypes were used as the reference group, the inverse of the RRRs and its confidence limits were taken.	
Study design	Head to head or single arm studies	Cohort or case control studies	Not stated	Not stated	(i) Case–control/ case-cohort/ cohort studies in which MD in cases, defined by subtype, is compared to non-cases and (ii) case-only designs where age-adjusted MD in ER+ cases is compared to that in ER- cases.	Prospective studies
Limits (language and date)	English; from 2000	Language not stated: up to December 31, 2015	English; date not stated	Stated to be no restrictions (assume this means none for language); date to February 8, 2013	English; 5th June 2012	Language not stated; January 1, 2004, through January 1, 2008
Limitations		Overall ES from all 6 articles not calculated, because the number of articles related to Asian women was small and because the breast density index varied	Very little information on systematic review methods	The authors did not formally use a quality assessment tool; the results from this meta-analysis reflect univariable associations only, as	Differences in density assessment methods. Restricted to English-language publications and only found studies conducted in North America and Europe, in	The studies reviewed had various designs, populations, and methods of analysing data. Although breast density is a strong risk factor for breast

		across articles. The subgroup analysis could not include results that were not divided by menopausal status. The analysis of premenopausal women was insufficient for dose-response meta-regression (DRMR). The subjects included only women who were born and lived in Asia (women born in Asia but emigrated overseas excluded). In the case-control studies, the most recent mammogram before breast cancer diagnosis were used, but this does not reflect the fact that breast density changes with age.		individual studies did not adjust their results for potential confounders, such as lesion size or histologic breast cancer subtype, thus precluding solid causal inference.	predominantly Caucasian women, thus other countries and ethnic groups, particularly at lower breast cancer risk are not included. Additionally, there was the lack of power to analyse combinations of ER and PR status.	cancer, BI-RADS has only modest reproducibility and more reproducible quantitative approaches are not validated or feasible for clinical use; so increased predictive accuracy may not be applicable to current clinical practice.

Methodological quality of included studies

Systematic reviews were assessed for quality using the AMSTAR criteria, which have been validated as a means to assess the methodological quality of systematic reviews and include establishing the research question and inclusion criteria before the conduct of the review, data extraction by at least two independent data extractors, comprehensive literature review with searching of at least two databases, key word identification, expert consultation and limits applied, detailed list of included/excluded studies and study characteristics, quality assessment of included studies and consideration of quality assessments in analysis and conclusions, appropriate assessment of homogeneity, assessment of publication bias and a statement of any conflict of interest. AMSTAR is not designed to generate an overall score. The quality appraisal is shown in Figure 10 below.

Figure 10. Quality appraisal for included studies for question 2b



Smith 2011[63] recommends tabulating the results of the systematic reviews, including the primary outcome of interest and the quality assessment (see Appendix 6). None of the studies stated that grey literature was included; none included a list of both included and excluded studies; none reported that the scientific quality of the included studies was assessed or used appropriately in formulating conclusions. Analyses were mainly narrative, which was appropriate.

Analysis of the evidence

Visual methods

Antoni 2013[68] focused on mammographic breast density as a risk factor by cancer type (estrogen receptor positive [ER+] and negative [ER-]), and found 19 studies, of which only seven provided analyses adjusted only for age, and of these, three used BIRADS and one used percent density. The review reported that mammographic density is a strong marker of breast cancer risk. For the eligible study using percent density, the relative risk of ER+ tumours was 1.38 (1.22, 1.57, $p<0.05$) for low vs. minimal density and the relative risk of ER- tumours was 0.95 (0.67, 1.34, not significant). These risks were not shown for the eligible BIRADS studies.

Bae 2016[65] investigated the association between mammographic breast density and breast cancer risk in Asian women using summary effect sizes (sES based on adjusted odds ratios [factors adjusted for not reported]) and found six studies (including three using percent density and one using Volpara [see below]). An overall ES reflecting information from all 6 articles was not calculated, because the number of articles was small and the breast density index varied across articles. For premenopausal women assessed using percent density, the sES was 3.23 (95% CI 2.23, 4.66; two studies). For postmenopausal women assessed using percent density, the sES was 1.62 (95% CI 1.13, 2.32; three studies). The authors concluded that breast cancer risk in Asian women increased with mammographic breast density measured using percent density.

Cummings 2009[64] (an update of McCormack 2006[69]) reviewed prospective studies about models and sex hormone levels to assess breast cancer risk, including one study assessing mammographic breast density using BI-RADS and four measuring percent density, in addition to the studies included in McCormack 2006[69]. All were adjusted for age; some studies adjusted for additional factors which were not stated except to say that studies that further adjust for body mass index or weight led to somewhat stronger associations. The authors found that breast density was strongly associated with breast cancer: relative risk vs. BIRADS category I was 2.03 (95% CI 1.61, 2.56) for BIRADS II; 2.95 (95% CI 2.32, 3.73) for BIRADS III; and 4.03 (95% CI 3.10, 5.26) for BIRADS IV. For measurement of percent density, vs. <5% dense area, the RR was 1.74 (95% CI 1.50, 2.03) for 5 – 24% density; 2.15 (95% CI 1.87, 2.48) for 25 – 49% density; 2.92 (95% CI 2.55, 3.34) for 50 – 74% density; and 4.20 (95% CI 3.61, 4.89) for >75% density.

Elias 2014[67] focused mainly on human epidermal growth factor receptor type 2 (HER2) overexpression (a marker of breast cancer aggressiveness), and found 14 studies which provided unadjusted results. The review reported that extremely dense breasts on mammography increased the chance of HER2 over-expression (BI-RADS breast density category 4 extremely dense had a pooled odds ratio of 1.37 for HER2 over-expression vs. BIRADS 1, 2 and 3; 95% CI 1.07–1.76, p=0.01; 9 studies), i.e. were associated with more aggressive cancers.

Huo 2014[66] found 37 studies including four providing results only adjusted for age: two using BIRADS, and two using (semi-automated) methods (see below). One of the BIRADS studies was reported as showing the OR of an interval cancer for women with dense breasts was 1.62, and the age-adjusted rate ratio was 2.45 for breast cancer incidence (no 95% CI shown). The other BIRADS study was reported as showing that BIRADS IV breasts were more often mammographically occult (no data shown).

Semi-automated methods

Huo 2014[66] found one study using Cumulus and reported that ≥50% density was associated with a 2.63-fold risk of developing breast cancer compared to density <10%; and high density was also associated with ER-positive tumours. The other study of a computer-assisted (semi-automated) method (not stated which) showed that dense area was a better predictor of breast cancer risk than percent density (but no data shown).

Automated methods

Bae 2016[65] reported for pre- and post-menopausal women assessed using Volpara, the summary effect size (sES) was 2.52 (95% CI 1.84, 3.46; one study).

3.2.4 Discussion

Seven studies were included in question 2a. All the studies found a reduced sensitivity of mammography and/or an increased risk of interval cancers with increasing mammographic breast density, in screening programmes in non-UK countries which have a shorter screening interval. Of the five systematic reviews we included in question 2b, the one with the most recent search date included Asian women only;[65] the previous one contained very limited information on systematic review methods so scored poorly on the AMSTAR criteria;[66] the one prior to that focused mainly on HER2 over-expression;[67] the one before that focused on cancer type (e.g. estrogen receptor positivity).[68] Cummings 2009[64] was an update of McCormack 2006[69] but did not report the population covered or other details of the included or excluded studies. In spite of these limitations, overall, the strength of the association between mammographic breast density and risk of breast cancer and the consistency of results between studies using varying methods, designs and locations suggests that mammographic breast density is an independent risk factor for breast cancer.

3.2.5 Summary

Question 2: NSC criterion 1: There should be robust evidence about the association between the risk or disease marker and serious or treatable disease: **Met.**

The evidence for the association between density and breast cancer was met for all density measurement methods.

3.3 Key question 3

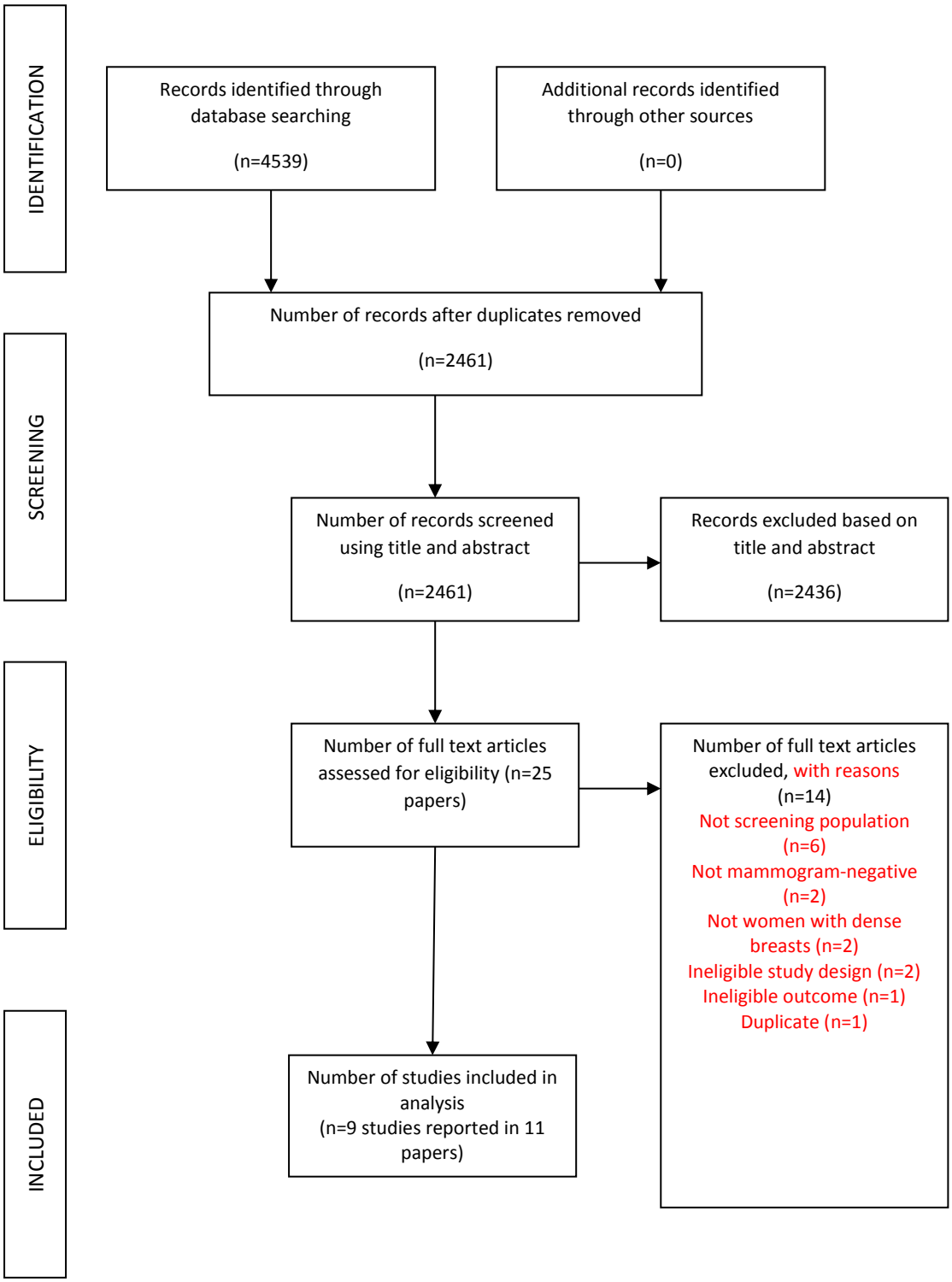
Question 3: What is the test accuracy of ultrasound following mammography in comparison to mammography to detect cancer in women with dense breasts?

This relates to NSC criterion 4:
"There should be a simple, safe, precise and validated screening test."

3.3.1 Description of the evidence

Searches of electronic databases identified 4539 unique studies. 258 records were examined at title and abstract stage, of which 25 were examined as full texts. Eleven of the papers (reporting on nine studies)[70-78] were subsequently included in the review, and 14 studies were excluded (listed in Appendix 3). The numbers of studies are shown in the PRISMA flow chart below (Figure 11).

Figure 11. PRISMA flow diagram for question 3



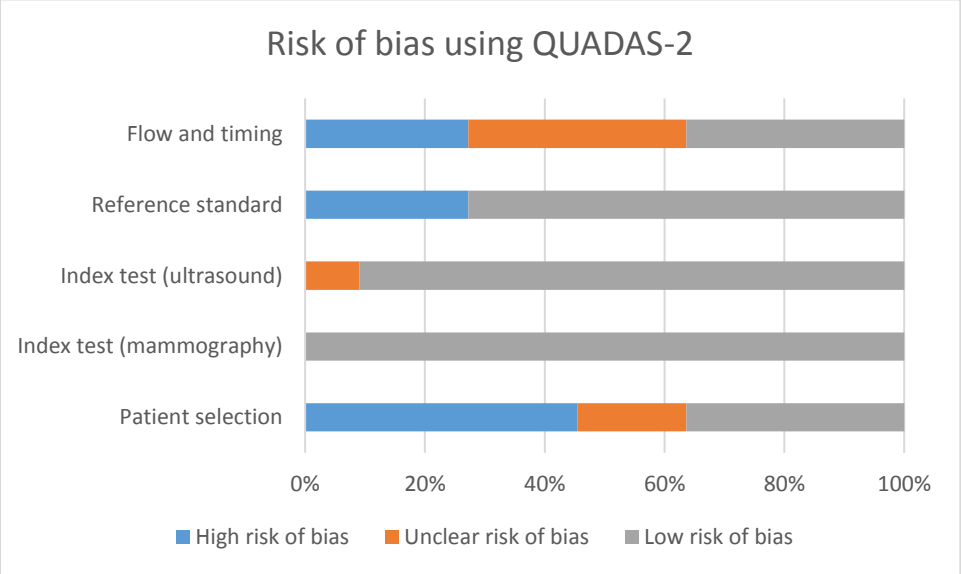
3.3.2 Characteristics of the included studies

During this update review, we found eleven papers reporting on nine studies, but none were classified as good-quality. Sample sizes ranged from 394[74] to 10,282,[77] and the studies were conducted in Italy,[76] Korea,[70,72,73,75] Sweden[78] and the USA.[71,74,77] Ages ranged from 24 or younger to at least 88 years, although some studies did not report the ages of the included women.

3.3.3 Methodological quality of included studies

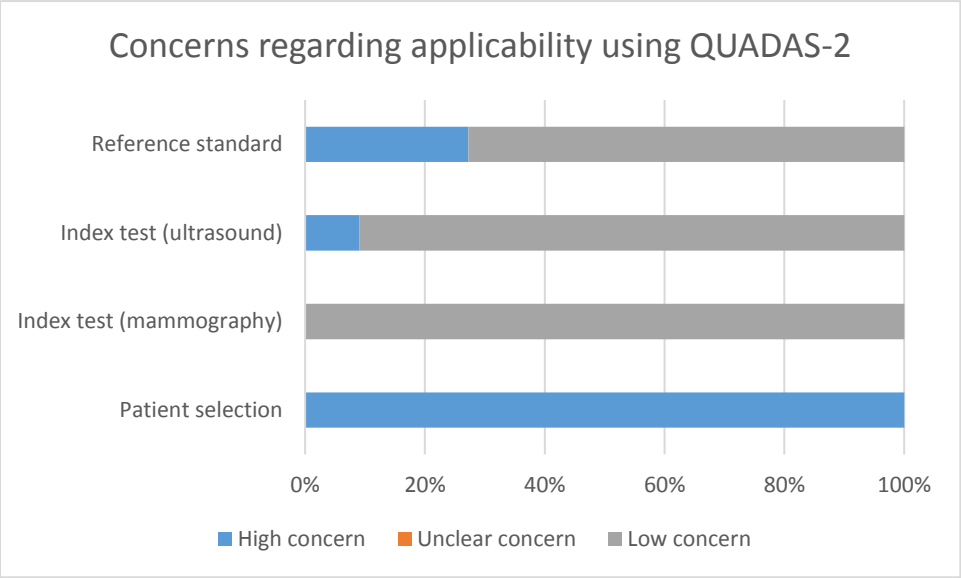
Including the two additional eligible studies from the USPTF review (Brem 2015[79] and Giuliano 2013[80]), quality appraisal was conducted on eleven studies. The adjusted QUADAS-2 quality assessment tool was used which provided two sets of data: firstly, the risk of bias and secondly, concerns regarding eligibility, which are shown in Figures 12 and 13, respectively. Patient selection was at high risk of bias in five (45%) studies[70-72,74,81] due to patients self-selecting whether or not to undergo ultrasound, and only a minority of patients took up the offer. There was a low risk of bias for the index tests (mammography or ultrasound) for all the studies except one (9%) study[73] in which the interpretation of the ultrasound used the non-standard "downgrade criteria". Three (27%) studies[71,76,81] did not follow women up for interval cancers, leading to a high-risk of bias for the reference standard and the flow/timing domains. In addition, the interval between the tests was unclear in four (36%) studies,[70,72,74,78] leading to an unclear risk of bias in the flow/timing domain.

Figure 12. Risk of bias for studies included in question 3 using QUADAS-2



All the studies were assessed as high concern regarding applicability due to differing populations not generalisable to the UK screening population (the proportion of women outside the 50-70 year age range was between 33%[80] and 60%[78] in seven studies; the other four did not report this percentage, but of these, one[75] was in Korea; in two,[74,81] only around 30% of eligible women participated, and in the other,[71] 67% of participants had risk factors compared with 26% in the overall screening population). There was a low concern about applicability for the index tests (mammography or ultrasound) for all the studies except one (9%) study[73] in which the interpretation of the ultrasound used the non-standard "downgrade criteria".

Figure 13. Concern regarding applicability for studies included in question 3



3.3.4 Analysis of the evidence

The USPTF[25] performed a systematic review of the test performance and clinical outcomes of supplemental screening with breast ultrasonography in women with dense breasts and negative mammography results. MEDLINE, PubMed, EMBASE, and Cochrane databases were searched from January 2000 to July 2015. This review found two good-quality studies (see Table 10 below) which reported that sensitivity of ultrasonography for women with negative mammography results ranged from 80% to 83%; specificity, from 86% to 95%; and positive predictive value (PPV) from 3% to 8%. Rates of additional cancer detection with ultrasonography were 4.4 per 1000 examinations (89% to 93% invasive); recall rates were 14%. The review reported that good-quality evidence was sparse. Studies were small and CIs were wide. Definitions of recall were absent or inconsistent. The review concluded that supplemental screening of women with dense breasts finds additional breast cancer but increases false-positive results. It is important to assess whether these results are generalisable to the UK population. The ultrasound studies in the USPTF review were examined to assess whether they would meet our inclusion criteria individually (see Table 10). We sought to identify whether the studies provided estimates of sensitivity, specificity, recall rates, biopsy rates, PPV and cancer detection rates of supplemental ultrasound which could be analysed alongside the data from the studies in our update review (see below). The results of our review may differ from the USPTF review because they included, and we excluded, studies of high-risk women, women outside of the population-based screening program, mixed screening and diagnostic populations, and film mammography; we also required data from women with dense breasts to be shown separately, which they did not.

Table 10. Papers in the USPTF review: sensitivity, specificity and eligibility for the update

A: USPTF review papers			**Eligible for our update review?**
Study	**Sensitivity (all patients in study)**	**Specificity (all patients in study)**	**Eligible for our review (and reason if not eligible)**
Berg 2012*	83%	86%	No – high risk women

Brancato, 2007	Not reported	Not reported	No – patients were self-referring to mammography, i.e., outside of the population-based screening program offered to women of 50-69 years.
Brem 2015[79]	Not reported	Not reported	Yes
Corsetti 2011*	80%	95%	No – film mammography not digital
Girardi 2013	Not reported	Not reported	No – women with dense breasts not shown separately
Giuliano 2013[80]	Not reported	Not reported	Yes
Hooley 2012 (	100%	77%	No – mixed screening and diagnostic population
Kelly 2010	68%	92%	No – high risk women
Leong 2012	100%	79%	No – film mammography not digital
Parris 2013	Not reported	Not reported	No –women with dense breasts not shown separately
Venturini 2013	Not reported	Not reported	No – women with dense breasts not shown separately
Weigert 2012[81]	Not reported	Not reported	Yes
Youk 2011	100%	72%	No - film mammography not digital

* Assessed as good quality in the USPTF review

Only Brem 2015[79] and Giuliano 2013[80] were included in our update data as separate studies; Weigert 2012[81] is an earlier publication from the same study as Weigert 2015[77] and Weigert 2017[82] which is included in our update. The Tables below show the eligible studies from the USPTF review (Table 11) and from our update searches (Table 12). We include the following information: quality issues, and whether studies provided evidence on sensitivity, specificity, recall rate, biopsy rate, PPV (of recall or of biopsy) and cancer detection rate of supplemental ultrasound in women with mammogram-negative dense breasts.

Table 11. Studies in the USPTF 2016 review: quality issues, and sensitivity, specificity, recall rate, biopsy rate, positive predictive value (of recall or of biopsy) and cancer detection rate of supplemental ultrasound in women with mammogram-negative dense breasts

USPTF review papers	**If eligible for our update review, data in women with mammogram-negative dense breasts only**								
Study	Quality issues	Sensitivity (%)	Specificity (%)	Recall rate (per 1000)	Biopsy rate (per 1000)	Positive predictive value of recall (%) = PPV_1	Positive predictive value of biopsy (%) = PPV_2	Benign biopsies (false positives) per 1000	Cancer detection rate (per 1000)
Brem 2015[79] (ABUS)	40.2% aged <50 yr, plus 6.7% >70 yr	Not reported	Not reported	2407/13107 = 184/1000	552/13107 = 42/1000	30/2407 = 1.2%	30/552 = 5.4%	522/13107 = 39.8/1000	30/13107 = 2.3/1000
Giuliano 2013 (ABUS)[80]	22.9% <50 yr plus 12.0% ≥70 yr	42/43 = 97.67%	3365/3375 = 99.70%	Not reported	52/3418 = 15.2/1000	Not reported	42/52 = 80.8%	10/3418 = 2.9/1000	42/3418 = 12.3/1000
Weigert 2012 (HHUS)[81]	Only 30% of eligible women had US. No follow up for interval cancers.	Not reported	Not reported	1196/8647 = 138/1000	418/8647 = 48.3/1000	28/1196 = 2.3%	28/418 = 6.7%	390/8647 = 45/1000	28/8647 = 3.2/1000

ABUS = automated ultrasound; HHUS = handheld ultrasound

Table 12. Studies from our update searches: quality issues, and sensitivity, specificity, recall rate, biopsy rate, positive predictive value (of recall or of biopsy) and cancer detection rate of supplemental ultrasound in women with mammogram-negative dense breasts

Update review papers	**Quality issues**	**Sensitivity (%)**	**Specificity (%)**	**Recall rate (per 1000)**	**Biopsy rate (per 1000)**	**Positive predictive value of recall (%) = PPV_1**	**Positive predictive value of biopsy (%) = PPV_2**	**Benign biopsies (false positives) per 1000**	**Cancer detection rate (per 1000)**
Chang 2015[70] (HHUS)	Median 47 (range 27-79) yr, i.e. >50% aged <50 yr	5/5 = 100%	624/985 = 63.4%	366/990 = 370/1000	Not reported	5/366 = 1.4%	Not reported	Not reported	5/990 = 5.1/1000
Destounis 2015[71] and Destounis	Patients self-selected for US after notification of dense breasts. Only 5.9% of those eligible participated. 17.93% aged <46	Not reported	Not reported	135/5434 = 248/1000 screens	100/4898 women = 20.4/1000	18/135 = 13.3%	18/100 = 18%	82/5434 = 15/1000	18/5434 = 3.3 per

2017[83] (HHUS)	yr; 4.27% >76 yr. No follow up for interval cancers								1000 screens
Hwang 2015[72] (HHUS)	25.3% of women with negative mammograms underwent US (women who requested US, regardless of risk factors, not only women with dense breasts). 12.5% of these lost to follow up. Median age 49.5 yr; range 30–76 yrs. 6.2% in their 30's, 44.2% in their 40's, 40.1% in their 50's, 8.3% in their 60's and 1.2% in their 70's.	8/9 = 88.9%	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported	Not reported
Kim 2016[73] (HHUS)	Mean ± SD: 51.2 ± 7.7 yr, range 24–78 yr, i.e. around 44% <50 yr and around 1% >70 yr. The focus of the study was on using "downgrade criteria" which would not be used in routine screening practice elsewhere.	9/9 = 100%	2340/3162 = 74%	831/3171 = 262/1000	147/3171 = 46/1000	9/831 = 1.1%	9/131 = 6.9%	122/3171 = 38/1000	9/3171 = 2.8/1000
Klevos 2017[74] (HHUS)	Only 32.5% of eligible women participated; small sample size	Not reported	Not reported	69/394 = 175/1000	26/394 = 66/1000	Not reported	Not reported	Not reported	0/394 = 0/1000
Moon 2015[75] (HHUS)	Self-selected for US; only 51.5% eligible participated. Mean 53.8 (range 40 to 87) yr	3/3 = 100%	1064/1653 = 64.4%	592/1656 = 357/1000	86/1656 = 52/1000	3/592 = 0.51%	2/86 = 2.33%	84/1656 = 51/1000	3/1656 = 1.8/1000
Tagliafico 2016[76] (HHUS)	Median 51 yr (IQR 44-78 yr; range, 38-88 yr). Not followed for interval cancers	-	-	88/3231 = 27/1000	47/3231 = 14.5/1000	23/88 = 26.1%	23/47 = 48.9%	24/3231 = 7.4/1000	23/3231 = 7.1/1000
Weigert 2015[77] and Weigert 2017[82] (HHUS)	Self-selected for US; only around 30% of eligible women participated. No follow up for interval cancers	-	-	1310/10282 = 127/1000	435/10282 = 42/1000	24/1310 = 1.8%	24/435 = 5.5%	411/10282 = 40/1000	24/10282 = 2.3/1000
Wilczek 2016[78] (ABUS)	Mean (SD) 49.5 (7.9), range 40-69 yr, i.e. >50% were <50 yr. Unclear how many patients did not consent to study and if those who consented were representative	4/9 = 44.4%	1625/1636 = 99.3%	15/1645 = 9.1/1000	12/1645 = 7.3/1000	4/15 = 26.7%	4/12 = 33.3%	8/1645 = 4.9/1000	4/1645 = 2.4/1000

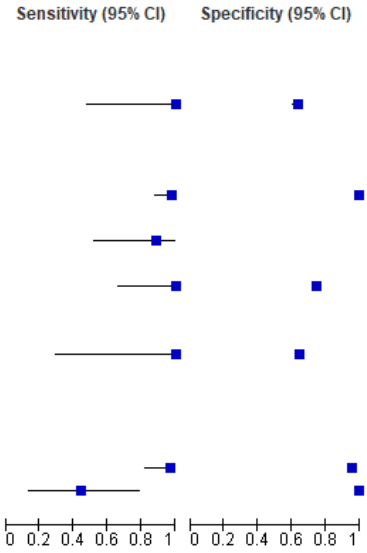
ABUS = automated ultrasound; HHUS = handheld ultrasound

Sensitivity and specificity

Including the data from the eligible USPTF studies and our update studies, the sensitivity of ultrasonography for women with dense breasts with negative mammography ranged from 44%[78] to 100%[70,73,75] (available data from seven studies) and specificity from 63%[70] to 100%[80] (available data from six studies; see Figure 14 below). The study with the highest values for both sensitivity and specificity[80] included around 35% of women outside the 50-70-year age range, so may not be generalisable to the UK screening population. Most of the studies had wide confidence intervals around the estimate of the sensitivity due to small numbers of events (the sum of the true positives [TP] plus false negatives [FN] was less than 10 people in five[70,72,73,75,78] of the seven studies providing data on sensitivity).

Figure 14: Forest plot of sensitivity and specificity of additional ultrasound in mammogram-negative dense breasts

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Berg 2012 (high risk)	0	0	0	0	Not estimable	Not estimable
Brancato 2007 (not screening pop)	0	0	0	0	Not estimable	Not estimable
Brem 2015 (no data)	0	0	0	0	Not estimable	Not estimable
Chang 2015	5	361	0	624	1.00 [0.48, 1.00]	0.63 [0.60, 0.66]
Corsetti 2011 (film)	0	0	0	0	Not estimable	Not estimable
Destounis (no data)	0	0	0	0	Not estimable	Not estimable
Girardi 2013 (not data)	0	0	0	0	Not estimable	Not estimable
Giuliano 2013	42	10	1	3365	0.98 [0.88, 1.00]	1.00 [0.99, 1.00]
Hooley 2012 (screen/diag)	0	0	0	0	Not estimable	Not estimable
Hwang 2015	8	0	1	0	0.89 [0.52, 1.00]	Not estimable
Kelly 2010 (high risk)	0	0	0	0	Not estimable	Not estimable
Kim 2016	9	822	0	2340	1.00 [0.66, 1.00]	0.74 [0.72, 0.76]
Klevos 2017 (no data)	0	0	0	0	Not estimable	Not estimable
Leong 2012 (film)	0	0	0	0	Not estimable	Not estimable
Moon 2015	3	589	0	1064	1.00 [0.29, 1.00]	0.64 [0.62, 0.67]
Parris 2013 (not dens)	0	0	0	0	Not estimable	Not estimable
Tagliafico 2016 (no data)	0	0	0	0	Not estimable	Not estimable
Venturini 2013 (not dens)	0	0	0	0	Not estimable	Not estimable
Weigert 15/17 (no data)	0	0	0	0	Not estimable	Not estimable
Weigert 2012	28	401	1	7450	0.97 [0.82, 1.00]	0.95 [0.94, 0.95]
Wilczek 2016	4	11	5	1625	0.44 [0.14, 0.79]	0.99 [0.99, 1.00]
Youk 2011 (film)	0	0	0	0	Not estimable	Not estimable



Recall rates and positive predictive value of recall

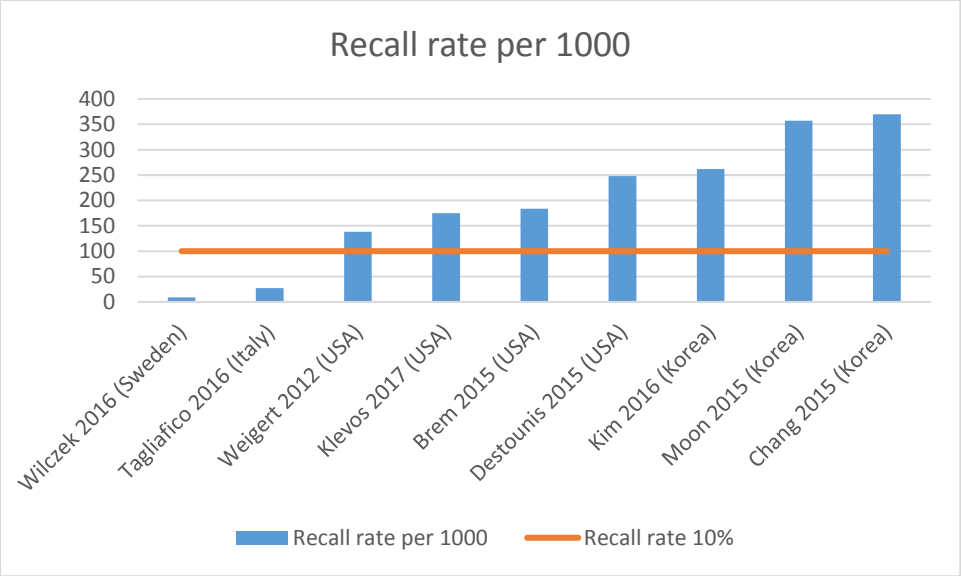
Including the data from the eligible USPTF 2016 studies and our update studies, recall rates were 9.1 per 1000[78] to 370 per 1000.[70] Quality assurance guidelines for breast cancer screening radiology from the NHS Breast Screening Programme[1] contain the following radiological quality standards (Table 13):

Table 13. Quality standard for mammographic recall rates

Objective	Criteria	Minimum standard	Achievable standard
To minimise the number of women screened who are referred for further tests	The percentage of women who are referred for assessment	(a) Prevalent screen < 10% Incident screen < 7%	(a) Prevalent screen < 7% Incident screen < 5%

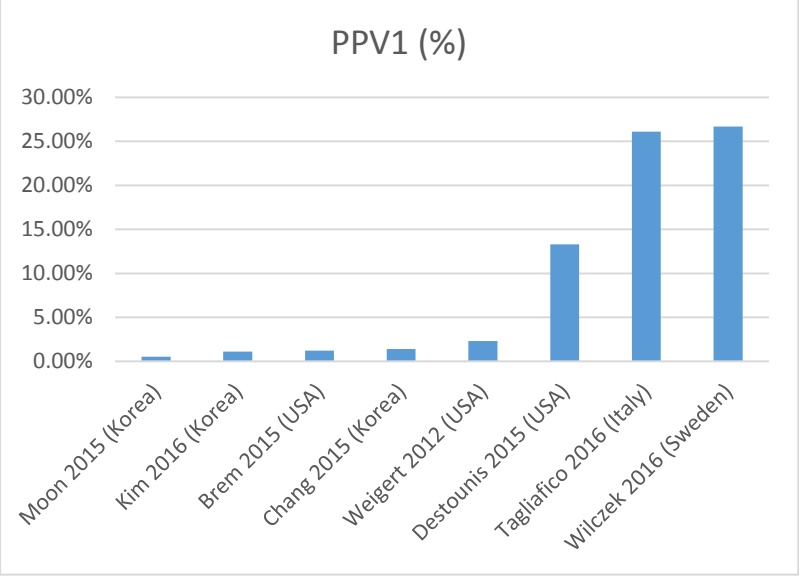
Of the ten studies providing data on recall rates, only two[76,78] had a recall rate for ultrasound of <10% (<100 per 1000); these two studies were conducted in Europe, in contrast to the other studies which were conducted in Korea or the USA, potentially reflecting differences in the patient populations and/or healthcare systems (see Figure 15).

Figure 15. Recall rates



The positive predictive value of recall (PPV_1; i.e. the likelihood of cancer among women who were recalled) ranged from 0.51%[75] to 26.7%;[78] higher (better) values were seen in the two European studies (see Figure 16).

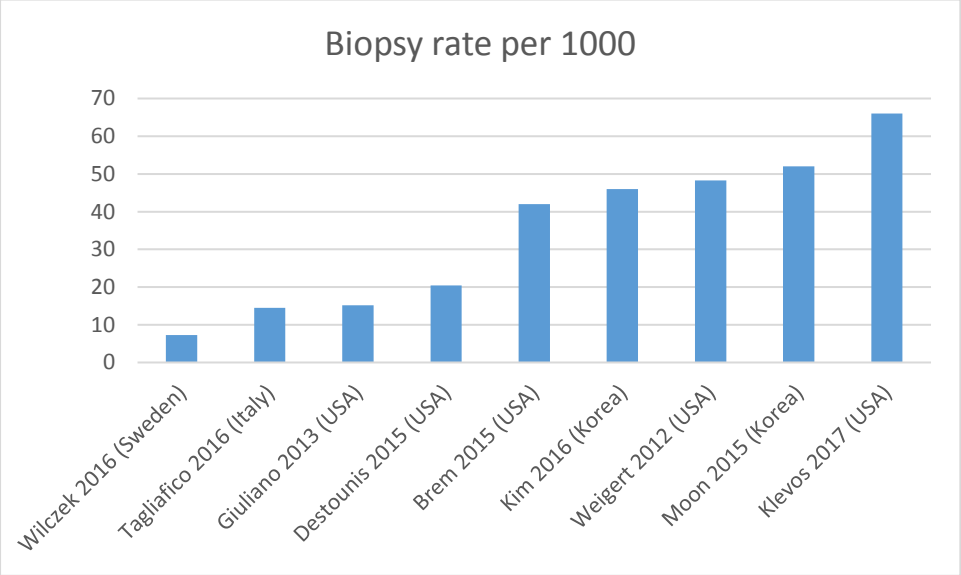
Figure 16. Positive predictive value of recall (PPV_1; i.e. the likelihood of cancer among women who were recalled)



Positive predictive value of biopsy and false positives

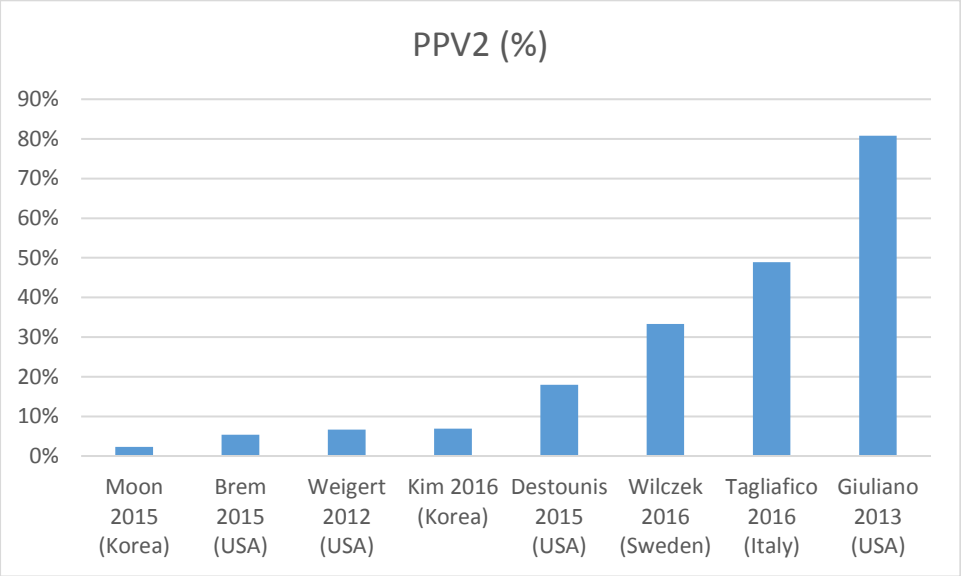
Including the data from the eligible USPTF studies and our update studies, biopsy rates were between 7.3 per 1000[78] and 66 per 1000;[74] the lowest rates were seen in the European studies (see Figure 17).

Figure 17. Biopsy rates per 1000



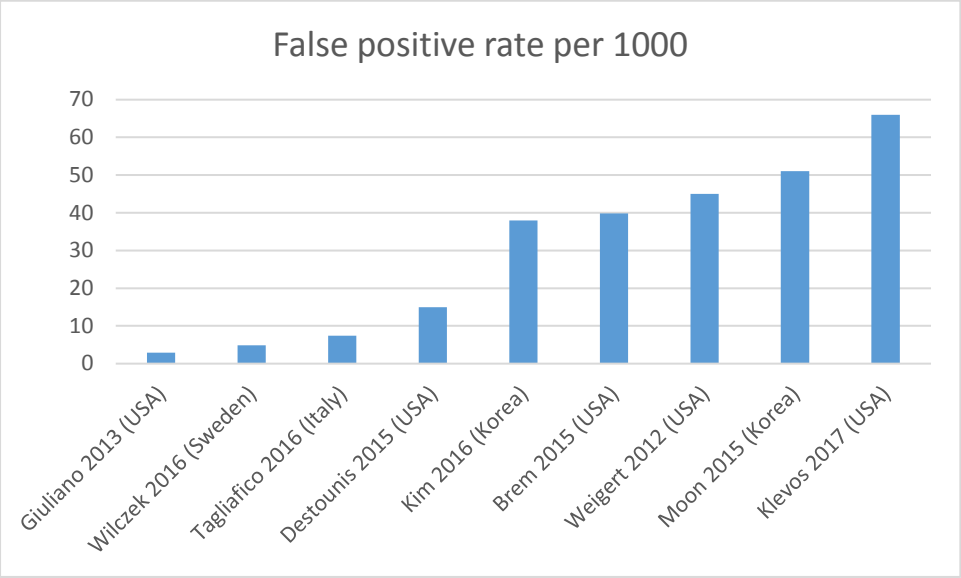
The positive predictive value of biopsy (PPV_2; i.e. the likelihood of cancer among women who had a biopsy) ranged from 2.33%[75] to 80.8%;[80] see Figure 18.

Figure 18. Positive predictive value of biopsy (PPV_2; i.e. the likelihood of cancer among women who had a biopsy)



Including the data from the eligible USPTF studies and our update studies, the rate of benign biopsies (false positives) ranged from 2.9 per 1000[80] to 51 per 1000;[75] see Figure 19.

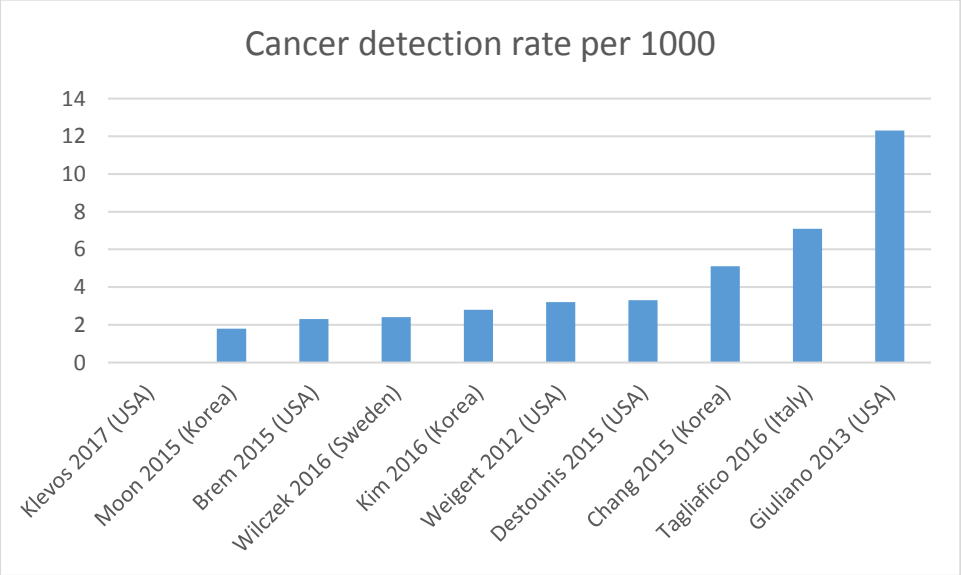
Figure 19. False positive rate per 1000



Cancer detection rates

The expected interval cancer rates after mammography are: 0–24 months: 1.2 invasive cancers per 1000 women screened; 25–36 months: 1.4 per 1000 women screened.[1] Rates of additional cancer detection with supplemental ultrasound were 0 per 1000[74] to 12.3 per 1000;[80] see Figure 20.

Figure 20. Cancer detection rates per 1000



Additional outcomes reported

Quality assurance guidelines for breast cancer screening radiology from the NHS Breast Screening Programme[1] (Table 14) state that one of the aims is to maximise the number of small invasive cancers detected (specifically invasive cancers < 15 mm in diameter).

Table 14. Aim of mammography is to maximise the number of small invasive cancers detected

Objective	Criteria	Minimum standard	Achievable standard
To maximise the number of small invasive cancers detected	The rate of invasive cancers < 15 mm in diameter detected in eligible women invited and screened	Prevalent screen ≥2 per 1000 Incident screen ≥2.3 per 1000	Prevalent screen ≥2.8 per 1000 Incident screen ≥3.1 per 1000

We therefore show the size of the cancers detected by supplemental ultrasound, as well as other features such as grade, lymph node involvement or distant metastases, and hormone receptor status, where these were reported. One study[71] reported that of the 100 BIRADS 4 or 5 lesions on ultrasound only that were biopsied/excised surgically, 18 (18%) were invasive cancers and the rest benign or atypical lesions. The invasive cancers comprised: invasive ductal carcinoma n=11 (61.11%); invasive lobular carcinoma n=5 (27.78%); invasive mammary carcinoma n=1 (5.56%) and metastatic carcinoma n=1 (5.56%). There were no DCIS. The invasive cancer grades were I: 5 (27.78%); II: 7 (38.89%); III: 4 (22.22%); and not specified: 2 (11.11%). The tumour sizes on sonography (cm) were: 0.1-0.5 cm: 1 (5.55%); 0.6-1.0 cm: 7 (38.89%); 1.1-1.5 cm: 4 (22.22%); 1.6-2.0 cm: 1 (16.67%); > 2.0 cm: 4 (16.67%) and not specified: 1 (5.55%). One patient did not undergo surgical excision because of extensive metastatic disease; of the 17 remaining patients, 4 (23.5%) had positive lymph nodes.

One study[72] reported 8 cancers detected by supplemental ultrasound only, of which 7 were invasive cancers (6 stage I; 1 stage II; 1 had positive lymph nodes) and 1 was DCIS (stage 0); they ranged in size from 0.5 cm to 2.4 cm (median, 0.9 cm) on ultrasound. Another study[73] reported that supplemental ultrasound screening detected 9 additional cancers, of which 7 were invasive cancers (3 invasive ductal carcinoma; 1 invasive lobular carcinoma; 1 mixed invasive ductal/lobular carcinoma; 1 invasive apocrine carcinoma and 1 mucinous carcinoma; 3 intermediate and 4 low grade) and 2 DCIS (low grade). The median size of the 9 cancers was 8 mm, ranging from 5 to 15 mm. None had lymph nodes or distant metastases; 7/9 (77.8%) were hormone receptor (HR) positive/HER2 negative and 2/9 (22.2%) were triple negative.

One study[76] reported that supplemental ultrasound screening detected an additional 23 cancers (17 invasive ductal carcinoma, 4 invasive lobular carcinoma, 1 mixed invasive [of which 3 grade 1, 10 grade 2, 5 grade 3 and 4 N/A] and 1 DCIS [low grade]). The mean tumour size was 15.1 mm (SD 4.8 mm); range 5 to 25 mm; 15 were ER+/PR+ or ER+/PR- or ER-/PR+; 2 ER-/PR- and 6 N/A; 7 had metastases in axillary nodes; 1 had micrometastases in axillary nodes; 13 were negative for lymph node involvement and 2 were N/A. HER2 status was 3+: 1; 2+: 0; 1+: 5; 0: 9 and 8 N/A. Another study[82] reported invasive ductal carcinoma with and without ductal carcinoma in situ: 14; invasive lobular carcinoma: 9; mixed type: 8; mucinous: 1; tubular: 1; ductal carcinoma in situ: 5; intracystic

or invasive papillary: 3; atypical ductal hyperplasia with papilloma: 3; lobular carcinoma in situ: 2. Of the 41 invasive cancers and DCIS, 9 were nuclear grade 1, 25 were nuclear grade 2, and 7 were nuclear grade 3; sizes ranged from 0.3 to 8.0 cm. 40 cancers had known hormonal status of which 33 were ER/PR+, 3 were ER+/PR-, one was ER-/PR+, one was ER/PR/HER+, and two were triple negative. Seven patients had positive metastatic lymph nodes. Four were in tumours that were nuclear grade 3 and were macro-metastatic and three were in tumours nuclear grade 2, one was macro-metastatic, and two were micro-metastatic. A final study[78] reported 4 additional screen-detected cancers with supplemental ultrasound: histological grades were: grade I: 2 (50.0%); grade II: 1 (25.0%); grade III: 1 (25.0%) and the mean (SD) size was 21.8 (12.6) mm, range 13 to 40 mm.

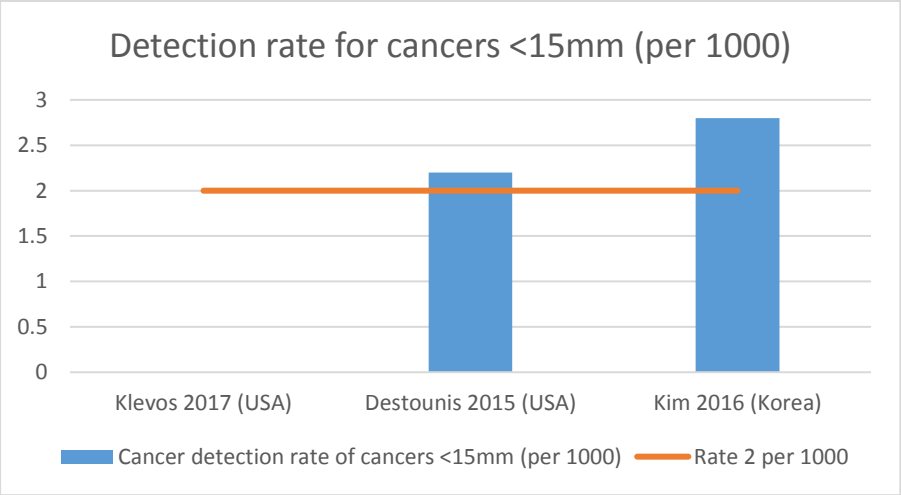
Table 15 and Figure 21 show the numbers of cancers of <15mm detected in the studies where this was reported.

Table 15: Numbers of cancers of <15mm detected

Study reference	**Overall cancer detection rate/1000**	**Cancers <15mm**	**Cancer detection rate/1000 for cancers <15mm**
[71]	18/5434 = 3.3 per 1000 screens	12	12/5434 = 2.2 per 1000*
[73]	9/3171 = 2.8/1000	9	9/3171 = 2.8 per 1000*
[74]	0/394 = 0/1000	0	0/394 = 0 per 1000*

* Calculated by us

Figure 21. Cancer detection rate for cancers <15mm (per 1000)



This suggests that some studies did detect a significant rate of small (<15mm) cancers, but there were only three studies[71,73,74] reporting the data to calculate such rates, of which one study[74] found no cancers at all.

3.3.5 Discussion

Study evidence

The results of our update review demonstrate that supplemental ultrasound can detect cancers that go undetected by mammography, including small (<15mm) cancers. Rates of additional cancer detection with supplemental ultrasound were 0 per 1000[74] to 12.3 per 1000;[80] and of small (<15mm) cancers were 0 per 1000[74] to 2.8 per 1000.[73] At least some of the cancers detected were of high grade and associated with positive lymph nodes. It is beneficial for mammography to detect small cancers, which without screening would present later as larger symptomatic cancers with a worse prognosis; mammography has been demonstrated to reduce the risk of mortality from breast cancer. However, it is unclear whether the additional detection by supplemental ultrasound of small, node-negative, low grade cancers (which have a good prognosis) would be beneficial in terms of reduction of mortality or reduction in the rate of interval cancers, as these lesions may represent overdiagnosis of cancers that would otherwise be found anyway at a later mammography screening round.

The sensitivity of ultrasonography for women with dense breasts with negative mammography ranged from 44%[78] to 100%[70,73,75] and specificity ranged from 63%[70] to 100%[80]. Recall rates were 9.1 per 1000[78] to 370 per 1000.[70] Of the ten studies providing data on recall rates, only two[76,78] (the European studies) had a recall rate for ultrasound of <10%. The positive predictive value of recall (PPV_1; i.e. the likelihood of cancer among women who were recalled) ranged from 0.51%[75] to 26.7%.[78] Biopsy rates were between 7.3 per 1000[78] and 66 per 1000;[74] the lowest rates were seen in the European studies. The positive predictive value of biopsy (PPV_2; i.e. the likelihood of cancer among women who had a biopsy) ranged from 2.33%[75] to 80.8%.[80] The rate of benign biopsies (false positives) ranged from 2.9 per 1000[80] to 51 per 1000.[75]

Study quality

The USPTF review found two good-quality studies but they did not meet our eligibility criteria, and one was using film mammography and the other involved high-risk women. Patient selection was at high risk of bias in five (45%) studies[70-72,74,81] due to patients self-selecting whether or not to undergo ultrasound, and only a minority of patients took up the offer. Three (27%) studies[71,76,81] did not follow women up for interval cancers, making it impossible to accurately assess the sensitivity of ultrasound. Most of the studies that did report sensitivity had wide confidence intervals around the estimate of the sensitivity due to small numbers of events (the sum of the true positives [TP] plus false negatives [FN] was less than 10 people in five[70,72,73,75,78] of the seven studies providing data on sensitivity).

Study applicability

Key issues in terms of the evidence base reviewed are its generalisability to the UK screening population. All the studies were assessed as high concern regarding applicability due to differing populations not generalisable to the general UK screening population (the proportion of women outside the 50-70 year age range was between 33%[80] and 60%[78] in seven studies; the other four did not report this percentage, but of these, one[75] was in Korea; in two,[74,81] only around 30% of eligible women participated, and in the other,[71] 67% of participants had risk factors compared with 26% in the overall screening population). In total, four studies were conducted in Korea,[70,72,73,75] three in the USA,[71,74,77] one in Italy[76] and one in Sweden.[78]

Consistency

Six of the seven studies with available data reported a sensitivity ≥89%; three of studies with available data reported the specificity below 75% and three above 75%. Recall and biopsy rates were lowest in the European studies,[76,78] with higher rates in the studies conducted in the USA or Korea.

3.3.6 Summary

Question 3: The NSC criterion 4: "There should be a simple, safe, precise and validated screening test": **Not met.**

Ultrasound can detect additional cancers among women with dense breasts and negative mammography, but estimates of sensitivity and specificity are uncertain as they are based on small numbers of events. The extra cancers detected come at the cost of high recall rates of between 9.1 to 370 per 1000, high biopsy rates of between 7.3 and 66 per 1000, and high benign biopsy rates (false positives) of between 2.9 to 51 per 1000. Variations between estimates may partly reflect the different populations and healthcare systems of the included studies. It is unclear to what extent the additional cancers represent overdiagnosis.

3.4 Key question 4 (cost-effectiveness)

Question 4. For women attending breast screening in the UK, what are the cost-consequences of adding mammographic density measurements, and then ultrasound for those found to have high mammographic breast density?

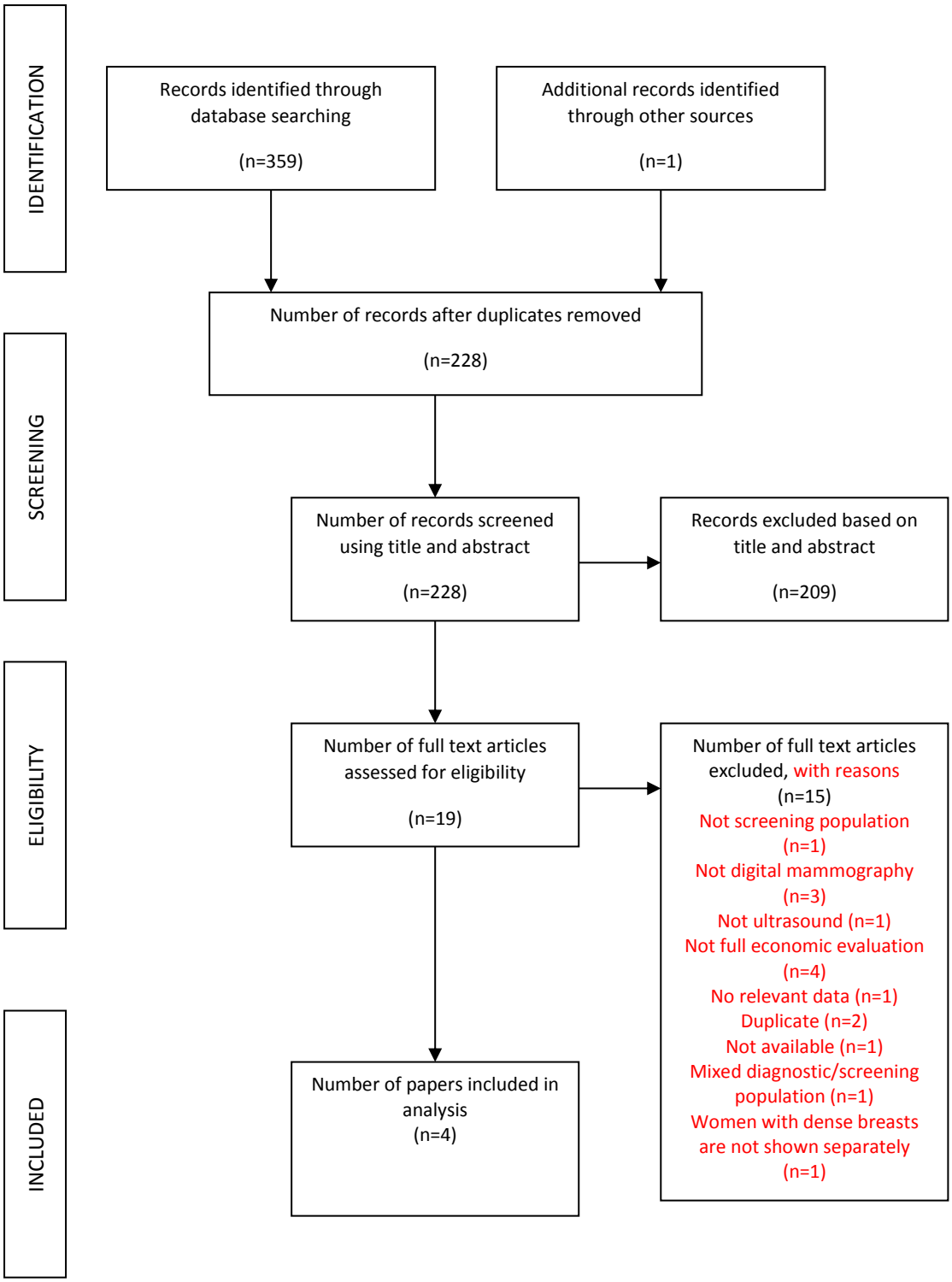
This relates to NSC criterion 14:

"The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (value for money). Assessment against this criteria should have regard to evidence from cost benefit and/or cost-effectiveness analyses and have regard to the effective use of available resource."

3.4.1 Description of the evidence

Figure 22 provides the PRISMA flow diagram for the cost-effectiveness question. We identified 228 unique records. Nineteen records were examined as full texts. Fifteen studies were excluded at full text stage; these are listed with the reason for exclusion in Appendix 3. This left four papers; one conducted in the UK[84] and three in the USA.[80,81,85]

Figure 22. PRISMA diagram for question 4



3.4.2 Characteristics of included studies

The included studies are described in Table 16.

Table 16. Characteristics of cost-effectiveness studies

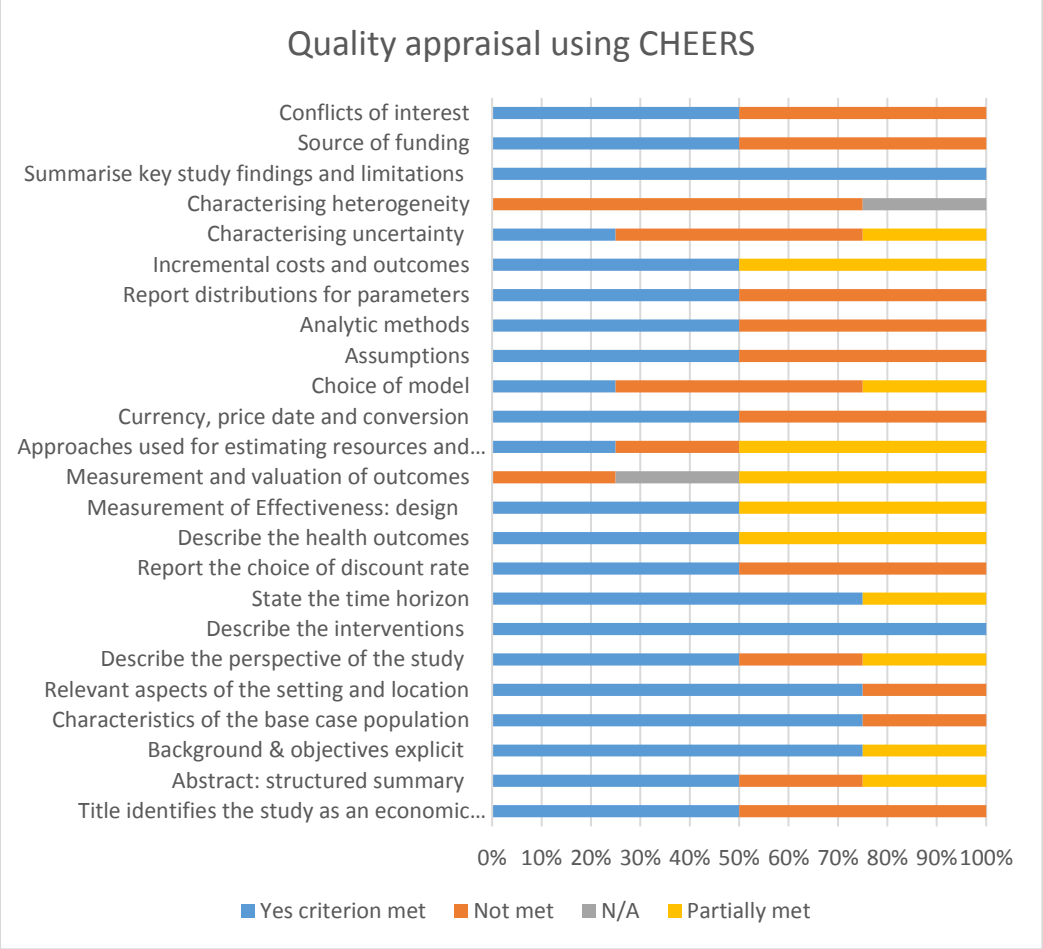
Author (Year)	Type of economic evaluation & model	Population studied	Comparators	Methods (perspective, time horizon and discount rate)	Methods (costs, outcomes, ICER and sensitivity analyses)
Giuliano 2013[80]	**EE:** CCA **Model:** None – but simple theoretical calculations	Women with dense breasts in a large screening population in the United States.	**Intervention:** Mammography plus ultrasound **Comparator:** Mammography only	**Study perspective:** Medicare and Medicaid reimbursement **Time horizon:** 1 year **Discount rate:** Not undertaken **Currency/price year:** US$, year not stated	**Outcomes:** additional treatment for missed cancers **Costs:** breast ultrasound, missed cancers, treatments **ICER:** cost per additional treatment for missed cancers **Sensitivity analyses:** Not undertaken
Gray 2017[84] (NB intervention also includes MRI)	**EE**: CUA **Model:** Decision-analytic model (discrete event simulation)	Women eligible for a national breast screening program (NBSP) in the UK	**Intervention:** Four approaches to stratified NBSP Risk 1 Risk 2 Masking - current screening approach with supplemental ultrasound offered to women with high breast density. Women with both high breast density and high risk of breast cancer were offered supplemental magnetic resonance imaging (MRI) instead of ultrasound	**Perspective:** National health Service **Time horizon:** Lifetime **Discount rate:** 3.5% for both costs and benefits **Currency/price year:** UK £ in 2015 prices	**Outcomes:** QALYs **Costs:** mammography, follow-up, biopsy, treatments, ultrasound, MRI **ICER:** cost per QALY gained **Sensitivity analyses:** One-way and probabilistic sensitivity analyses

			Risk 1 with masking **Comparator:** Current UK NBSP and no screening		
Sprague 2015[85]	**EE**: CEA **Model:** 3 micro-simulation models	Women eligible for breast screening in USA. Biennial screening for 50-74 year olds; Annual screening for 40-74 year olds.	**Intervention:** Mammography plus supplemental ultrasound **Comparator:** Mammography alone	**Perspective:** Federal Payer **Time horizon:** Lifetime **Discount rate:** 3% for both costs and benefits **Currency/price year:** US $ in 2013 prices	**Outcomes:** QALYs **Costs:** mammography screening, ultrasound, additional imaging, biopsy, cancer treatment **ICER:** cost per QALY gained **Sensitivity analyses:** One-way sensitivity analyses
Weigert 2012[81]	**EE:** CCA **Model:** None	Women with normal mammograms but dense breasts in the USA	**Intervention:** Mammography plus ultrasound **Comparator:** Mammography alone	**Perspective:** Not stated **Time horizon:** 1 year **Discount rate:** Not undertaken **Currency/price year:** US$, year not stated	**Outcomes:** Number of breast cancers detected **Costs:** average reimbursement by CPT-code and insurance company relating to mammograms, ultrasounds and biopsy's including staff time. **ICER:** Cost per breast cancer found **Sensitivity analyses:** Not undertaken

3.4.3 Methodological quality of included studies

All the studies described fully the interventions, findings and their limitations. Three studies reported adequately the objectives,[80,84,85] the time horizon,[81,84,85] setting/location[81,84,85] and aspects of the population studied.[80,84,85] Only two[84,85] reported fully the perspective of the study, discount rate, health outcomes used in the analysis, currency and price year for reporting costs, any assumptions made with the analysis, analytic methods used for the reporting the results, results reported as incremental costs and outcomes, the source of funding; whilst in the other two studies[80,81] these were reported partially or not at all (see Figure 23).

Figure 23. Quality assessment of included studies for question 4



3.4.4 Analysis of the evidence

A recent cost-utility study[84] conducted in the UK found that the current screening approach plus supplemental ultrasound offered to women with high mammographic breast density (defined using VDG3 and VDG4), with ultrasound and MRI for women at high risk, does not appear to be a cost-effective alternative when compared with the current UK National Breast Screening Programme (NBSP):

- ICER vs. No screening (3.5% DR): £30,772 per QALY gained

- ICER vs. UK NBSP (3.5% DR): £212,947 per QALY gained
- ICER vs. No screening (1.5% health, 3.5% costs DR): £15,065 per QALY gained
- ICER vs. UK NBSP (1.5% health, 3.5% costs DR): £105,412 per QALY gained.

As this was the only UK study, it was analysed in depth (see Table 17).

Table 17. Analysis of the UK cost-effectiveness study

Reference	Gray 2017[84]
Interventions and comparators	**Interventions** **Risk 1**: a risk-based stratification defined by the risk algorithm plus density and texture measures. Three strata (with associated screening intervals) were defined by 10-y risks of breast cancer of 1) <3.5% (3-yearly), 2) 3.5%–8% (2-yearly), and 3) >8% (annually) **Risk 2**: a risk-based stratification defined by the same algorithm as risk 1 but with strata defined by dividing the population into thirds on the basis of 10-y risk (tertiles): 1) the lowest risk tertile (3-yearly), 2) the middle tertile (2-yearly), and 3) the highest risk tertile (annually) **Masking** (covering up of tumors in mammograms by dense breast tissue): current screening approach with supplemental ultrasound offered to women with high breast density, defined using Volpara density grade 3 or 4. High risk was defined as >8% 10-y risk of breast cancer. Women with both high breast density and high risk of breast cancer were offered supplemental magnetic resonance imaging instead of ultrasound. **Risk 1 with masking**: the risk 1 stratification approach together with the strategy described in the masking approach **Comparators** **Current UK NBSP**: women between 50 and 70 y with screening every 3y using mammography **No screening**: no use of mammography in the population for screening purposes; all cancers would present with clinical signs or symptoms
Research question	To identify the incremental costs and consequences of stratified national breast screening programs (stratified NBSPs) and key drivers of relative cost-effectiveness.
Study type	Cost-effectiveness analysis
Study population	Women eligible for an NBSP. Mean +/- SD age (y): base case 48.93 +/- 1.09
Institutional setting	National health care service (NHS)
Country/currency	United Kingdom/£. National currency (£) at 2014 prices
Funding source	Part of a European collaborative project called Adapting Breast Cancer Screening Strategy Using Personalised Risk Estimation (ASSURE). The ASSURE project was funded from a collaborative project grant within the FP7-HEALTH-2012- INNOVATION-1 call (project number: 306088).
Analytical perspective	NHS
Effectiveness parameters	Multiple data sources were used: systematic reviews of effectiveness and utility and cohort studies embedded in existing NBSPs. Mammography and ultrasound sensitivity/specificity etc, interval cancers, survival and effectiveness of MRI referenced. Mammography • Sensitivity by tumor size modelled as logistic-type function • β1: sets increase with size 1.47 • β2: sets sensitivity relative to size 6.51

	• Maximum sensitivity 0.95% • Sensitivity by VDG, used to calculate relative sensitivity given tumor size • Sensitivity VDG1 85.0% • Sensitivity VDG2 77.6% • Sensitivity VDG3 69.0% • Sensitivity VDG4 58.6% • Recall rate 4.0 per 100 examinations • False-positive biopsy proportion 2.4% • Proportion of screen-detected cancers that are DCIS 20.3% • Clinically detected (interval cancers) • Cancer size at clinical detection, mean 6.5 doublings (22.62mm) • Cancer size at clinical detection, SD 0.535 doublings • Survival after breast cancer diagnosis • γ NPI 1 -5.413 • γ NPI 2 -4.023 • γ NPI 3 -2.465 • γ Advanced cancer, age <50 y -0.527 • γ Advanced cancer, age 50–69 y -0.537 • γ Advanced cancer, age ≥70 y -0.849 US cancer detection • VDG3/4 incremental cancers detected with supplemental US 3 per 1000 examinations • False-positive (recall) rate, US 98 per 1000 examinations • Biopsy rate, US 0.4% Assumed same as mammography • Proportion cancers detected by supplemental US that are DCIS 21% Assumed same as mammography MRI cancer detection • VDG3/4 incremental cancers detected with supplemental US 5 per1000 examinations • False-positive (recall) rate, MRI 41.15 per 1000 examinations • Biopsy rate, MRI 3.03% • Proportion of cancers detected by supplemental MRI that are DCIS 14.3%
Intervention costs	Multiple data sources were used: published studies reporting costs, and cohort studies embedded in existing NBSPs.

	Cost data referenced plus expert opinion. Costs • Mammography £54 • Follow-up, mean £95 • Biopsy, mean £160 • NPI 1 treatment, mean £11,630 • NPI 2 treatment, mean £12,978 • NPI 3 treatment, mean £15,405 • Advanced cancer, mean £23,449 • Screening ABUS £80 • Screening HHUS £80 • Screening MRI £220 • Stratification process £10.57
Indirect costs	Costs to individual women were excluded from the analysis
Health-state valuations/utilities	Multiple data sources were used: systematic reviews of effectiveness and utility, and cohort studies embedded in existing NBSPs. Utilities referenced Utility • Early breast cancer, first year 0.696 • Early breast cancer, subsequent years 0.779 • Advanced breast cancer, first year 0.685 • Advanced breast cancer, subsequent years 0.685
Modelling	A decision-analytic model (discrete event simulation). A *de novo* model was developed. The conceptualisation process identified that the model required three components to represent: the stratification approach, breast cancer natural history with screening, and the diagnosis and treatment process after a cancer detected by screening. A discrete event simulation (DES) model was used to represent these three components.
Transition probabilities for model	Extensive definitions of various parameters/equations used; also referenced to supplementary material
Time horizon	Lifetime

Discount rates applied in the model for costs and outcomes	3.5% for both costs and benefits (base case) 3.5% for costs and 1.5% for benefits (sensitivity analysis)
Results/analysis: Measure of benefit reported	QALYs
Clinical outcome/benefits estimated for each intervention/strategy	Screening program — QALYs (3.5% discount rate) — Cost (£,2015; 3.5% DR) No screening — 17.6919 — 246 Current UK NBSP — 17.7095 — 654 Risk 1 — 17.7119 — 694 Risk 2 — 17.7181 — 858 Masking — 17.7102 — 809 Risk 1 and masking — 17.7124 — 870
Synthesis of costs and benefits	Screening program ICER vs. **No screening** (3.5% DR) **UK NBSP** (3.5% DR) **No screening** (1.5% health, 3.5% costs) **UK NBSP** (1.5% health, 3.5% costs) No screening — NA — NA — NA — NA Current UK NBSP — £23,197 — NA — £11,343 — NA Risk 1 — £22,413 — £16,689 — £11,363 — £11,565 Risk 2 — £23,435 — £23,924 — £11,425 — £11,592 Masking — £30,772 — £212,947 — £15,065 — £105,412 Risk 1 and masking — £30,532 — £75,254 — £14,707 — £33,199 DR = discount rate Masking and risk 1 and masking were dominated by the next alternative (current NBSP and risk 1 stratified NBSP, respectively). The ICERs for the remaining comparisons were £23,197 per QALY for the current NBSP compared with no screening, £16,689 per QALY for risk 1 stratified NBSP compared with masking, and £26,749 for risk 2 stratified NBSP compared with masking and risk 1 stratified NBSP. The risk 1 and risk 2 stratified NBSPs were relatively cost-effective when compared with the current UK NBSP. The masking stratified NBSP does not appear to be a cost-effective alternative when compared with the current UK NBSP. When compared with no screening, all screening programs may be considered cost-effective.
Statistical analysis	Not shown

Sensitivity analysis	One-way sensitivity analyses were used to explore the impact of selected input parameters (referenced to supplementary material). Probabilistic sensitivity analysis (PSA) was performed to quantify the effect of the joint uncertainty.
Scenarios tested in sensitivity analysis	Input parameters and discount rates were varied
Results of the sensitivity analysis	Using an alternative discounting rate of 3.5% for costs and 1.5% for benefits resulted in relatively lower estimated incremental cost-effectiveness ratios (ICERs) for all stratified NBSPs compared with the UK NBSP. One-way sensitivity analysis showed that the reported total costs, total QALYs, and ICERs were sensitive to natural history parameter values ($\alpha 2$ and mean tumour size at clinical detection) and screening performance of mammography ($\beta 2$). ICERs for stratified programs were moderately sensitive to the cost of stratification although costs would need to be several times the base-case value for ICERs to increase beyond a threshold of £30,000 per QALY. In all alternative programs, total costs were sensitive to the treatment cost parameters; varying these parameters, however, did not greatly change the ICERs compared with the base case. Estimates of total QALYs were sensitive to the utility weights for cancer states; varying utility weights moderately altered the ICERs of stratified programs compared with the NBSP. The results were relatively insensitive (within the ranges tested) to the probability of recall, costs of MRI, the relative sensitivity of mammography by VDG group, and US/MRI additional cancer detection rate.
Conclusions/implications	A risk stratified NBSP is potentially a cost-effective use of health care resources when compared with the current UK NBSP.
Implications of the evaluation for practice	This early model-based cost-effectiveness analysis provides indicative evidence for decision makers to understand the key drivers of costs and QALYs for exemplar stratified NBSP. Key drivers of cost-effectiveness were discount rate, natural history model parameters, mammographic sensitivity, and biopsy rates for recalled cases. A key assumption was that the risk model used in the stratification process was perfectly calibrated to the population.

The first study in the USA[81] used a cost-consequence analysis and reported that using costs of $250 (approximate £ equivalent at 22 February 2018: £179) per ultrasound and $2,400 (approximate £ equivalent £1,719) per ultrasound-guided biopsy, the cost per breast cancer found was estimated to be $110,241 (approximate £ equivalent £78,940). However, they reported few details of their assumptions and analytical methods. The second study in the USA[80] used theoretical calculations and found that the cost differential for additional treatment between Stage 1 and Stage 2 breast cancer was $10,467 (approximate £ equivalent £7,495). They also reported that the cost-benefit of early detection of stage 1 disease results in annual capital cost savings of $22.75 (approximate £ equivalent £16.29) per screened patient in the USA population, according to their model. However, they did not present details of their assumptions or analytical model, or any actual or derived data to support improved breast cancer mortality with the addition of ultrasound. The third study in the USA[85] (which met the majority of the CHEERS quality criteria) used three micro-simulation models and the authors reported that supplemental ultrasound screening for women with dense breasts undergoing screening mammography would substantially increase costs while producing relatively small benefits in terms of breast cancer deaths averted and QALYs gained. The ICER was $325,000 (approximate £ equivalent £232,723) per QALY gained for women with heterogeneously or extremely dense breasts (biennial screening). Restricting supplemental ultrasound screening to women with extremely dense breasts the ICER was $246,000 (approximate £ equivalent £176,153) per QALY gained (biennial screening). For annual screening the ICERs were even higher than biennial screening.

3.4.5 Discussion

Study evidence

Only the UK study[84] was designed as a cost-effectiveness analysis; the authors collected and reported the required information for an economic evaluation, and concluded that supplemental screening was not cost-effective. The USA study[85] meeting the majority of the CHEERS criteria reported that supplemental ultrasound screening for women with dense breasts undergoing screening mammography would substantially increase costs while producing relatively small benefits in terms of breast cancer deaths averted and QALYs gained. The other two studies from the USA[80,81] provided insufficient details to fully evaluate their findings.

Study quality

On the CHEERS checklist, the UK study[84] met 16 of the 24 quality criteria, and one of the studies conducted in the USA[85] met 17 of the 24 criteria. The other two studies conducted in the USA met only four[80] and five[81] of the 24 criteria.

Study applicability

The intervention in the UK study[84] included not only ultrasound screening for women with dense breasts but also MRI screening for women at high risk, so the cost-effectiveness of the ultrasound component only cannot be properly established. The other three studies[80,81,85] reflect the healthcare system in the USA.

Consistency

The two studies[84,85] meeting the majority of the CHEERS criteria both suggest that supplemental ultrasound is not cost-effective.

3.4.6 Summary

Question 4: NSC criterion 14. “The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (value for money). Assessment against this criterion should have regard to evidence from cost-benefit and/or cost-effectiveness analyses and have regard to the effective use of available resource”: **Not met.**

There is insufficient evidence for cost-effectiveness of supplemental ultrasound, and the available evidence suggests that it is not currently cost-effective.

Section 4: Discussion

4.1 Evidence and assessment of NSC screening criteria

We examined five key questions relating to ultrasound as an add-on test after negative mammography screening in women with dense breasts:

1. What are the reliability and concordance of available methods to measure mammographic breast density? (NSC criterion 4)

2a. Is mammographic breast density a risk factor for cancers being missed during screening (masking on mammograms/false negatives/interval cancers)? (NSC criterion 1)

2b. Is mammographic breast density a risk factor for developing breast cancer? (NSC criterion 1)

3. What is the test accuracy of ultrasound following mammography in comparison to mammography to detect cancer in women with dense breasts? (NSC criterion 4)

4. For women attending breast screening in the UK, what are the cost-consequences of adding mammographic density measurements, and then ultrasound for those found to have high mammographic breast density? (NSC criterion 14)

For key question 1, even allowing for the expected changes in density over time, we found wide variation in density assessment within and between readers for visual methods. Semi-automated methods are more consistently reliable than visual methods in research settings, but similar high inter-reader reliability values may not be reproduced in clinical screening practice. With automated volumetric mammographic breast density measurements, a more consistent density assessment of serial screening mammograms was observed than with the density assessment performed by trained clinicians. However, automated methods such as Volpara and Quantra differ from each other; concordance between methods is not generally high as they define density in different ways and there is no gold standard applicable to all breast density measurements. While MRI has been suggested as a type of gold standard, discrepancies occur between breast density measurement methods and this gold standard, particularly at higher densities.[10]

For key question 2a all the studies found a reduced sensitivity of mammography and/or an increased risk of interval cancers with increasing mammographic breast density. Of the systematic reviews we included in question 2b, the strength of the association between mammographic breast density and risk of breast cancer and the consistency of results between studies using varying methods, designs and locations suggests that mammographic breast density is an independent risk factor for breast cancer.

For key question 3 we found that ultrasound can detect additional cancers among women with dense breasts and negative mammography (rates of additional cancer detection with ultrasound were 0 per 1000 to 7.1 per 1000, and of small [<15mm] cancers were 0 per 1000 to 2.8 per 1000). At least some of the cancers detected were of high grade and associated with positive lymph nodes. It is beneficial for mammography to detect small cancers, which without screening would present later as larger symptomatic cancers with a worse prognosis; mammography has been demonstrated to

reduce the risk of mortality from breast cancer. However, it is unclear whether the additional detection by supplemental ultrasound of small, node-negative, low grade cancers (which have a good prognosis) would be beneficial in terms of reduction of mortality or reduction in the rate of interval cancers, as these lesions may represent overdiagnosis of cancers that would otherwise be found anyway at a later mammography screening round. Sensitivity of additional ultrasound ranged from 44% to 100% and specificity from 63% to 100%. The extra cancers detected came at the cost of high recall rates of between 9.1 to 370 per 1000 (only 20% of the studies providing data on recall rates had a recall rate for ultrasound below 10%). The positive predictive value of recall (PPV_1) ranged from 0.51% to 26.7%. Biopsy rates were between 7.3 and 66 per 1000, and the positive predictive value of biopsy (PPV_2) ranged from 2.33% to 80.8%. The rate of benign biopsies (false positives) ranged from 2.9 to 51 per 1000.

For key question 4 we found only 4 eligible papers; one conducted in the UK and three in the USA. Only the UK study was designed as a cost-effectiveness analysis, but the intervention in that study included not only ultrasound screening for women with dense breasts but also MRI screening for women at high-risk, so the cost-effectiveness of the ultrasound component alone cannot be properly established. There is insufficient evidence for cost-effectiveness of supplemental ultrasound, and the available evidence suggests that it is not currently cost-effective.

NSC criterion	Our questions addressing this criterion	Met/ not met?	Key reasons
Criterion 1: There should be robust evidence about the association between the risk or disease marker and serious or treatable disease	Question 2a. Is mammographic breast density a risk factor for cancers being missed during screening (masking on mammograms/ false negatives/ interval cancers)? Question 2b. Is mammographic breast density a risk factor for developing breast cancer?	Met	Strong consistent association between mammographic breast density and risk of breast cancer. Consistent finding of reduced sensitivity of mammography and/or increased risk of interval cancers with increasing mammographic breast density.
Criterion 4: There should be a simple, safe, precise and validated screening test	Question 1: What are the reliability and concordance of available methods to measure mammographic breast density? Question 3. What is the test accuracy of ultrasound following mammography in comparison to	Not met	It is difficult to validate the density methods when there is no gold standard applicable to all breast density measurements, concordance between methods is low, and even automated methods are not interchangeable. Ultrasound is not precise because it leads to large numbers of false positives,

	mammography to detect cancer in women with dense breasts?		and while it can detect additional cancers not found on mammography, we do not have evidence on whether this reduces interval cancers in the screening programme or mortality, or to what extent this represents overdiagnosis.
Criterion 14: The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (value for money). Assessment against this criterion should have regard to evidence from cost-benefit and/or cost-effectiveness analyses and have regard to the effective use of available resource	Question 4. For women attending breast screening in the UK, what are the cost-consequences of adding mammographic density measurements, and then ultrasound for those found to have high mammographic breast density?	Not met	There is insufficient evidence for cost-effectiveness of supplemental ultrasound, and the available evidence suggests that it is not currently cost-effective.

Although not systematically investigated in this review, some evidence relating to other NSC was identified, including:

Criterion 5. The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.

The data here relate to women with heterogeneously or extremely dense breasts (BIRADS categories 3 and 4), whereas if a cut-off level were chosen only including women with extremely dense breasts (BIRADS 4), different values would be obtained, e.g. for sensitivity of ultrasound. Estimating cost-effectiveness at different density thresholds might be practical and worthwhile.

Criterion 9. There should be an effective intervention for patients identified through screening, with evidence that intervention at a pre-symptomatic phase leads to better outcomes for the screened individual compared with usual care.

Data are currently lacking on the benefit to the individual of earlier intervention after mammographic breast density assessment and ultrasound screening, as the proportion of cases which are reducing interval cancers or overdiagnosis is not known.

Criterion 11. There should be evidence from high quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity.

There is no RCT evidence of supplemental ultrasound reducing mortality, and such studies might not be realistic. However, RCTs might be justifiable examining reductions in morbidity (interval cancers) using mammographic breast density assessment and supplemental ultrasound with a longer follow up and more screening rounds.

Criterion 13. The benefit gained by individuals from the screening programme should outweigh any harms, for example from overdiagnosis, overtreatment, false positives, false reassurance, uncertain findings and complications.

It is unclear whether the benefits outweigh the harms, particularly due to the high rate of false positives, and the possibility of overdiagnosis and overtreatment.

Criterion 18. Adequate staffing and facilities for testing, diagnosis, treatment and programme management should be available prior to the commencement of the screening programme.

Introducing density assessment and supplemental ultrasound would require additional facilities in terms of personnel and equipment for screening, and there would also be an effect on the number of biopsy samples requiring laboratory processing. Visual density assessment methods show a strong relationship between density and cancer, despite inter-observer variability, but may be impractical for population-based screening; automated methods are likely to be more practical for risk stratification. Logistical challenges could include the inherent risk of increasing the complexity of the screening pathway by separating off a cohort of women for additional tests, and the need to update the National Breast Screening Service (NBSS) system to record density data. Of note, the American College of Radiology recently (November 2017) updated its statement[86] on the reporting of breast density in mammography reports and patient summaries, which now includes the following: "Supplemental screening should be a thoughtful choice after a complete risk assessment, not an automatic reaction to breast density itself."

Recent publications also suggest that automated breast density measures may contribute to risk stratification, and more accurate risk prediction could enable better targeting of risk-reducing interventions e.g. lifestyle modification.[21] For example, among women participating in the "Predicting Risk of Cancer at Screening" (PROCAS) study, Volpara density grades predicted subsequent cancer even after adjustment for other personal and familial risk factors (adjusted odds ratio 3.00, 955 CI 1.54 to 5.86 for Volpara density grade 4 versus grade 1).[21] Therefore density assessment may be valuable as part of a holistic risk assessment, rather than as an automatic gateway to supplemental ultrasound screening. Another recent publication compared Volpara and Quantra versus MRI, and found that while percent breast density can be accurately measured using automated volumetric software programs, values should not be used interchangeably between methods.[10] Other authors have noted that moving towards a standardised assessment of mammographic breast density for clinical applications would be hugely complex, and involve consideration of how consistent the method is across X-ray systems, modalities and over time, as well as how feasible the method is in terms of integration into health information technology systems and clinical practice.[20] The UK NHSBSP screens over two million women per year and

authors in the UK have noted that in order to be practicable, any breast composition risk marker would have to be fully-automated with minimal human resource implications.[87]

Other authors have recently concluded that most women with dense breasts and no other risk factors are likely to experience more harms than benefits with supplemental screening ultrasonography.[88] Other barriers to the wider use of ultrasound in screening might include the need for trained technologists or physicians to perform and interpret scans.[89] Particular issues are that every normal breast has a different and unique ultrasound appearance; there are no consistent and reproducible landmarks except the nipple, pectoralis muscle and axilla; and small or subtle cancers may blend in with fibrocystic changes; ultrasound therefore requires a highly skilled and experienced technologist.[89]

4.2 Strengths and limitations

We conducted a systematic review for each of the key questions. We searched four databases, date limits were applied, and only articles in the English language were included; therefore it is possible that relevant articles might have been missed by this strategy, although search terms were broad. We included a wide scope of questions including cost-effectiveness. We built on a recent review of the relevant literature and used a systematic approach to the design of our search strategies and to inclusion and exclusion and quality assessment. Sifting and data extraction were performed by two reviewers. We performed thorough quality appraisal in duplicate; no studies were excluded on grounds of quality.

An adequate number of studies were found for question 1 but we found no multi-centre studies that included representative samples of women and raters, plus tests within a 2-year time-frame. We did not include all methods of density measurement; we excluded older methods which have been superseded, however, other methods may predict cancer (e.g. visual analogue scales),[21] but these were not prioritised by the advisory group prior to finalising the protocol. A limitation of the quality assessment tool used for the studies in question 1 is that five of the eleven questions relate to blinding, with studies marked down for a lack of blinding, which may be important for research studies, but in real-world screening practice, readers would not be blinded to previous assessment of density or clinical information, and therefore real-world studies would be inappropriately graded as lower quality. Another limitation of research studies may be their design for readers to focus all their attention on breast density, making density the most important finding on the mammograms, which is not the case in real practice in which density is usually a secondary focus of attention.

It should be noted that our review was designed to apply to the general screening populations (which will include a proportion of high-risk women) but we excluded studies performed solely in high-risk women. The rationale for excluding papers on non-screening populations for question 1 (performance of the density measures during screening) was that there are reasons to believe that women in diagnostic/mixed population studies would not be representative of women who participate in screening (e.g. by distribution of breast density or age). We included 28 studies in the review for question 1; the largest one included 83 readers and mammograms from 87,066 women. These appear to give us a good sense of the performance of the density measures. However, diagnostic/mixed population studies could provide additional useful information about density

measures in general. And density screening with ultrasound may be a reasonable strategy as part of a programme of care for high-risk women.

An adequate number of studies were found for question 2. However, in question 2a, none of the studies we found were at low-risk of bias. Question 2b was covered by several systematic reviews; however, they covered limited populations (Asian women only) or focused on cancer subtypes (HER2 over-expression or estrogen receptor positivity), or did not report the population covered or other details of the included or excluded studies so scored poorly on AMSTAR quality criteria.
We did not duplicate the USPTF systematic review but we built on that work by conducting an update, using similar search terms and quality assessment tools. However, full details of these methods were not available so relied on interpretation of the information that was present in the report. We complemented this method by carrying out our own quality assessment using the QUADAS-2 tool on both our update papers and also the original papers included in the USPTF review. However, it should be noted that some of the papers included in the USPTF review did not match our inclusion criteria (e.g. they included film mammography as well as digital). There were no good-quality studies in the question 3 update to the USPTF review – the authors of that review also noted the poor quality of the evidence base.

We found only four studies eligible for question 4, including only one fully-published UK cost-effectiveness study.

4.3 Conclusion/general interpretation of the results in the context of other evidence, and implications for policy, practice and future research

There is strong and consistent evidence both that dense breasts increase the risk of breast cancer and decreases the sensitivity of mammography to detect cancers. Given that mammographic breast density is a risk factor for development of breast cancer (question 2b), and that breast cancer may be missed by mammography in women with dense breasts (question 2a), women with dense breasts may require supplementary screening over and above the mammography offered to women without this risk factor. For this to be feasible, it would require a) a reliable method of mammographic breast density assessment (question 1) and b) a supplementary test that was sensitive, specific, accurate (question 3) and cost-effective (question 4).

The studies included in question 1 found that overall, there is variation in density assessment within and between readers for visual assessment methods. Objective automated methods appear to be more reliable, although there is insufficient high-quality evidence to support this. Automated methods are not equivalent to each other. In question 3, we found that supplemental ultrasound can detect additional cancers in women with negative mammography and dense breasts, but at a cost of additional false-positives and unnecessary biopsies. Further it is not known if the additional cancers represents overdiagnosis. In question 4, we found that cost-effectiveness studies from the US and the UK concluded that supplementary ultrasound in all women with heterogeneously or

extremely dense breasts does not appear to be cost-effective. Focusing on women with extremely dense breasts only would be more cost-effective than including women with heterogeneously dense breasts also.

Implications for research

The implications for research include the need for:

- Assessment of methods of measuring mammographic breast density which offer consistency, reliability and validity within a general screening population, which have a proven strong relationship to both risk of cancer and risk of masking and which are practical in terms of scale up into the screening programme. This is required alongside
- stronger evidence for benefits in terms of reduction in interval cancers or breast cancer mortality from supplemental ultrasound after mammographic breast density assessment.
- A randomised controlled trial including cost-effectiveness assessment would provide the necessary answers to the question of whether density assessment followed by ultrasound for women with dense breasts would be clinically and cost effective within the screening programme. Follow up long enough to assess the different types of cancer found, along with any reductions in interval cancers, would be required in order to address the issue of potential overdiagnosis. However there are challenges to performing such a trial including "contamination" between clusters and potentially very high costs. In addition screening technology continues to evolve.[89]

Implications for practice

The implication for practice is that if density assessment followed by supplementary ultrasound screening were undertaken in the current NHS breast screening programme, women could be categorised differently between readers or screening occasions unless a standardized programme-wide method of density assessment were used. Such a programme however could lead to increased anxiety and resource use (for women identified as at higher risk who might not actually be at higher risk), and to confusion for women whose categorization changed. Our review suggests that the numbers of false positives and additional biopsies are unlikely to be justified, and that there is as yet no clear cost effectiveness evidence to balance the benefits, harms and costs.

Section 5: Conflict of interest and funding statement

Funding: NSC
The authors have no conflict of interest to declare.

The commissioners gave feedback on the study protocol but had no role in the collection, analysis or interpretation of data, or in the writing of the report.

Team members' contributions

The Division of Health Sciences is located within Warwick Medical School. Warwick Medical School brings together experts in clinical and cost effectiveness reviewing, medical statistics, health economics and modelling. All team members checked and agreed to the final version of the report. The team that carried out the work were:

Name: Dr Jacoby Patterson
Address: Division of Health Sciences, Warwick Medical School, Gibbet Hill, Coventry, CV4 7AL
Contribution: Protocol development, assessment for eligibility, quality assessment of studies, data extraction, and report writing

Name: Dr Chris Stinton
Address: Division of Health Sciences, Warwick Medical School, Gibbet Hill, Coventry, CV4 7AL
Contribution: Protocol development, assessment for eligibility, quality assessment of studies, data extraction, and report writing

Name: Dr Lena Alkhudairy
Address: Division of Health Sciences, Warwick Medical School, Gibbet Hill, Coventry, CV4 7AL
Contribution: Assessment for eligibility, quality assessment of studies, data extraction, commenting on the draft report and final version of the report

Name: Dr Amy Grove
Address: Division of Health Sciences, Warwick Medical School, Gibbet Hill, Coventry, CV4 7AL
Contribution: Assessment for eligibility, quality assessment of studies, data extraction, commenting on the draft report and final version of the report

Name: Dr Pam Royle
Address: Division of Health Sciences, Warwick Medical School, Gibbet Hill, Coventry, CV4 7AL
Contribution: Database searches and procurement of articles, commenting on the draft report and final version of the report

Name: Hannah Fraser
Address: Division of Health Sciences, Warwick Medical School, Gibbet Hill, Coventry, CV4 7AL
Contribution: Administration and liaison; data extraction checking and article procurement, commenting on the draft report and final version of the report

Name: Dr Hema Mistry
Address: Division of Health Sciences, Warwick Medical School, Gibbet Hill, Coventry, CV4 7AL
Contribution: Assessment for eligibility, quality assessment of studies, data extraction, commenting on the draft report and final version of the report

Name: Payagalage Senaratne
Address: Division of Health Sciences, Warwick Medical School, Gibbet Hill, Coventry, CV4 7AL
Contribution: Assessment for eligibility, quality assessment of studies, data extraction, commenting on the draft report and final version of the report

Name: Prof Aileen Clarke
Address: Division of Health Sciences, Warwick Medical School, Gibbet Hill, Coventry, CV4 7AL
Contribution: Overseeing project and report writing, commenting on the draft report and final version of the report

Name: Dr. Sian Taylor-Phillips
Address: Division of Health Sciences, Warwick Medical School, Gibbet Hill, Coventry, CV4 7AL
Contribution: Overseeing project and report writing

REFERENCES

1. NHS Breast Screening Programme. Quality Assurance Guidelines for Breast Cancer Screening Radiology. Publication No 59. 2011. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/470579/nhsbsp59_QA_radiology_uploaded_231015.pdf
2. Cancer Research UK. Breast cancer statistics: breast cancer incidence (invasive). 2017. http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer#heading-Zero.
3. Falcon S, Williams A, Weinfurtner J, Drukteinis J. Imaging Management of Breast Density, a Controversial Risk Factor for Breast Cancer. *Cancer Control* 2017; **24**(2): 125-36.
4. Brentnall AR, Harkness EF, Astley SM, et al. Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort. *Breast Cancer Research* 2015; **17**(1): 147.
5. Cancer Research UK. Breast cancer diagnosis and treatment statistics: Routes to diagnosis of breast cancer. 2016. http://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/diagnosis-and-treatment#heading-Seven.
6. Howell A, Astley S, Warwick J, et al. Prevention of breast cancer in the context of a national breast screening programme. *Journal of Internal Medicine* 2012; **271**(4): 321-30.
7. Wanders JO, Holland K, Veldhuis WB, et al. Volumetric breast density affects performance of digital screening mammography. *Breast Cancer Research & Treatment* 2017; **162**(1): 95-103.
8. Holmberg L, The Working Party for Higher-Risk Breast Screening. Report of the Working Party for Higher Risk Breast Screening 2015. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/442491/report-working-party-higher-risk-breast-screening.pdf (accessed.
9. Eng A, Gallant Z, Shepherd J, et al. Digital mammographic density and breast cancer risk: a case-control study of six alternative density assessment methods. *Breast Cancer Research* 2014; **16**(5): 439.
10. Rahbar K, Gubern-Merida A, Patrie JT, Harvey JA. Automated Volumetric Mammographic Breast Density Measurements May Underestimate Percent Breast Density for High-density Breasts. *Acad Radiol* 2017; **24**(12): 1561-9.
11. D'Orsi C, Sickles E, Mendelson E, Morris E. ACR BI-RADS Atlas: Breast Imaging Reporting and Data System. 5th ed. Reston, VA: American College of Radiology; 2013.
12. Irshad A, Leddy R, Ackerman S, et al. Effects of Changes in BI-RADS Density Assessment Guidelines (Fourth Versus Fifth Edition) on Breast Density Assessment: Intra- and Interreader Agreements and Density Distribution. *AJR American Journal of Roentgenology* 2016; **207**(6): 1366-71.
13. van der Waal D, den Heeten GJ, Pijnappel RM, et al. Comparing Visually Assessed BI-RADS Breast Density and Automated Volumetric Breast Density Software: A Cross-Sectional Study in a Breast Cancer Screening Setting. *PLoS ONE [Electronic Resource]* 2015; **10**(9): e0136667.
14. Jeffers AM, Sieh W, Lipson JA, et al. Breast Cancer Risk and Mammographic Density Assessed with Semiautomated and Fully Automated Methods and BI-RADS. *Radiology* 2017; **282**(2): 348-55.
15. Llobet R, Pollan M, Anton J, et al. Semi-automated and fully automated mammographic density measurement and breast cancer risk prediction. *Computer Methods & Programs in Biomedicine* 2014; **116**(2): 105-15.

16. Lobbes MB, Cleutjens JP, Lima Passos V, et al. Density is in the eye of the beholder: visual versus semi-automated assessment of breast density on standard mammograms. *Insights Into Imaging* 2012; **3**(1): 91-9.
17. Conant EF, Keller BM, Pantalone L, Gastounioti A, McDonald ES, Kontos D. Agreement between Breast Percentage Density Estimations from Standard-Dose versus Synthetic Digital Mammograms: Results from a Large Screening Cohort Using Automated Measures. *Radiology* 2017; **283**(3): 673-80.
18. Destounis S, Johnston L, Highnam R, Arieno A, Morgan R, Chan A. Using Volumetric Breast Density to Quantify the Potential Masking Risk of Mammographic Density. *AJR American Journal of Roentgenology* 2017; **208**(1): 222-7.
19. Ekpo EU, McEntee MF, Rickard M, et al. QuantraTM should be considered a tool for two-grade scale mammographic breast density classification. *British Journal of Radiology* 2016; **89**(1060): 20151057.
20. Destounis S, Arieno A, Morgan R, Roberts C, Chan A. Qualitative Versus Quantitative Mammographic Breast Density Assessment: Applications for the US and Abroad. *Diagnostics* 2017; **7**(2): 31.
21. Astley SM, Harkness EF, Sergeant JC, Warwick J, Stavrinos P. A comparison of five methods of measuring mammographic density: a case-control study. *Breast Cancer Research* 2018; **20**(10).
22. Sprague BL, Conant EF, Onega T, et al. Variation in Mammographic Breast Density Assessments Among Radiologists in Clinical Practice: A Multicenter Observational Study. *Annals of Internal Medicine* 2016; **165**(7): 457-64.
23. Burton A, Maskarinec G, Perez-Gomez B, et al. Mammographic density and ageing: A collaborative pooled analysis of cross-sectional data from 22 countries worldwide. *PLoS Med* 2017; **14**(6): e1002335.
24. Bailey S, Sigal B, Plevritis S. A Simulation Model Investigating the Impact of Tumor Volume Doubling Time and Mammographic Tumor Detectability on Screening Outcomes in Women Aged 40–49 Years. *J Natl Cancer Inst* 2010; **102**(16): 1263-71.
25. Melnikow J, Fenton JJ, Whitlock EP, et al. A Systematic Review for the U.S. Preventive Service Task Force U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews, Report No.: 14-05201-EF-3. *Agency for Healthcare Research and Quality (US)* 2016; **Supplemental Screening for Breast Cancer in Women With Dense Breasts**: A Systematic Review for the U.S. Preventive Service Task Force U.S. Preventive Services Task Force Evidence Syntheses, formerly Systematic Evidence Reviews, Report No.: 14-05201-EF-3.
26. American College of Radiology. ACR statement on reporting breast density in mammography reports and patient summaries. 2012. www.acr.org/About-Us/Media-Center/Position-Statements/Position-Statements-Folder/Statement-on-Reporting-Breast-Density-in-Mammography-Reports-and-Patient-Summaries.
27. Public Health England. Requirements for UK NSC evidence summaries. 2016. https://www.gov.uk/government/publications/uk-nsc-evidence-review-process/appendix-f-requirements-for-uk-nsc-evidence-summaries
28. Public Health England. Criteria for appraising the viability, effectiveness and appropriateness of a screening programme 2015. https://www.gov.uk/government/publications/evidence-review-criteria-national-screening-programmes/criteria-for-appraising-the-viability-effectiveness-and-appropriateness-of-a-screening-programme.
29. Lucas N, Macaskill P, Irwig L, et al. The reliability of a quality appraisal tool for studies of diagnostic reliability (QAREL). *BMC Medical Research Methodology* 2013; **13**: 111.

30. Hayden J, van der Windt D, Cartwright J, Cote P, Bombardier C. Assessing Bias in Studies of Prognostic Factors. *Ann Intern Med* 2013; **158**: 280-6.
31. Shea B, Hamela C, Wells G, et al. AMSTAR is a reliable and valid measurement tool to assess the methodological quality of systematic reviews. *J Clin Epidemiol* 2009; **62**(10): 1013-20.
32. Whiting P, Rutjes A, Westwood M, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. *Annals of Internal Medicine* 2011; **155**(8): 529-36.
33. Husereau D, Drummond M, Petrou S, et al. Consolidated Health Economic Evaluation Reporting Standards (CHEERS) statement. *BMJ* 2013; **346**: f1049.
34. Landis J, Koch G. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159-74.
35. Cicchetti D. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment* 1994; **6**(4): 284-90.
36. Ekpo EU, Mello-Thoms C, Rickard M, Brennan PC, McEntee MF. Breast density (BD) assessment with digital breast tomosynthesis (DBT): Agreement between QuantraTM and 5th edition BI-RADS. *Breast* 2016; **30**: 185-90.
37. Abdolell M, Tsuruda K, Schaller G, Caines J. Statistical evaluation of a fully automated mammographic breast density algorithm. *Computational & Mathematical Methods in Medicine* 2013; **2013**: 651091.
38. Singh T, Sharma M, Singla V, Khandelwal N. Breast Density Estimation with Fully Automated Volumetric Method: Comparison to Radiologists' Assessment by BI-RADS Categories. *Academic Radiology* 2016; **23**(1): 78-83.
39. Mazor RD, Savir A, Gheorghiu D, Weinstein Y, Abadi-Korek I, Shabshin N. The inter-observer variability of breast density scoring between mammography technologists and breast radiologists and its effect on the rate of adjuvant ultrasound. *European Journal of Radiology* 2016; **85**(5): 957-62.
40. Holland K, van Zelst J, den Heeten GJ, et al. Consistency of breast density categories in serial screening mammograms: A comparison between automated and human assessment. *Breast* 2016; **29**: 49-54.
41. Osteras BH, Martinsen AC, Brandal SH, et al. Classification of fatty and dense breast parenchyma: comparison of automatic volumetric density measurement and radiologists' classification and their inter-observer variation. *Acta Radiologica* 2016; **57**(10): 1178-85.
42. Gweon HM, Youk JH, Kim JA, Son EJ. Radiologist assessment of breast density by BI-RADS categories versus fully automated volumetric assessment. *AJR American Journal of Roentgenology* 2013; **201**(3): 692-7.
43. Kang E, Lee EJ, Jang M, et al. Reliability of Computer-Assisted Breast Density Estimation: Comparison of Interactive Thresholding, Semiautomated, and Fully Automated Methods. *AJR American Journal of Roentgenology* 2016; **207**(1): 126-34.
44. Seo JM, Ko ES, Han BK, Ko EY, Shin JH, Hahn SY. Automated volumetric breast density estimation: a comparison with visual assessment. *Clinical Radiology* 2013; **68**(7): 690-5.
45. Eom H, Cha J, Kang J, Choi W, Kim H, Go E. Comparison of variability in breast density assessment by BI-RADS category according to the level of experience. *Acta Radiologica* 2017.
46. Garrido-Estepa M, Ruiz-Perales F, Miranda J, et al. Evaluation of mammographic density patterns: reproducibility and concordance among scales. *BMC Cancer* 2010; **10**: 485.
47. Sartor H, Lang K, Rosso A, Borgquist S, Zackrisson S, Timberg P. Measuring mammographic density: comparing a fully automated volumetric assessment versus European radiologists' qualitative classification. *European Radiology* 2016; **26**(12): 4354-60.

48. Alshafeiy TI, Wadih A, Nicholson BT, et al. Comparison Between Digital and Synthetic 2D Mammograms in Breast Density Interpretation. *AJR American Journal of Roentgenology* 2017; **209**(1): W36-W41.
49. Harvey JA, Gard CC, Miglioretti DL, et al. Reported mammographic density: film-screen versus digital acquisition. *Radiology* 2013; **266**(3): 752-8.
50. Raza S, Mackesy MM, Winkler NS, Hurwitz S, Birdwell RL. Effect of Training on Qualitative Mammographic Density Assessment. *Journal of the American College of Radiology* 2016; **13**(3): 310-5.
51. Irshad A, Leddy R, Lewis M, et al. Changes in Breast Density Reporting Patterns of Radiologists After Publication of the 5th Edition BI-RADS Guidelines: A Single Institution Experience. *American Journal of Roentgenology* 2017; **209**: 943-8.
52. Kerlikowske K, Ma L, Scott C, et al. Combining quantitative and qualitative breast density measures to assess breast cancer risk. *Breast Cancer Research* 2017; **19**(1): 97.
53. Busana MC, Eng A, Denholm R, et al. Impact of type of full-field digital image on mammographic density assessment and breast cancer risk estimation: a case-control study. *Breast Cancer Research* 2016; **18**(1): 96.
54. Martinez Gomez I, Casals El Busto M, Anton Guirao J, Ruiz Perales F, Llobet Azpitarte R. Semiautomatic estimation of breast density with DM-Scan software. *Radiologia* 2014; **56**(5): 429-34.
55. Pollan M, Llobet R, Miranda-Garcia J, et al. Validation of DM-Scan, a computer-assisted tool to assess mammographic density in full-field digital mammograms. *Springerplus* 2013; **2**(1): 242.
56. Osteras BH, Martinsen AC, Brandal SH, et al. BI-RADS Density Classification From Areometric and Volumetric Automatic Breast Density Measurements. *Academic Radiology* 2016; **23**(4): 468-78.
57. Bédard M, Martin N, Krueger P, Brazil K. Assessing Reproducibility of Data Obtained With Instruments Based on Continuous Measurements. *Experimental aging research* 2000; **26**: 353-65.
58. Rawashdeh MA, Bourne RM, Ryan EA, et al. Quantitative measures confirm the inverse relationship between lesion spiculation and detection of breast masses. *Academic Radiology* 2013; **20**(5): 576-80.
59. Nelson HD, O'Meara ES, Kerlikowske K, Balch S, Miglioretti D. Factors Associated With Rates of False-Positive and False-Negative Results From Digital Mammography Screening: An Analysis of Registry Data. *Annals of Internal Medicine* 2016; **164**(4): 226-35.
60. Timmermans L, Bleyen L, Bacher K, et al. Screen-detected versus interval cancers: Effect of imaging modality and breast density in the Flemish Breast Cancer Screening Programme. *European Radiology* 2017; **13**: 13.
61. Holland K, van Gils CH, Mann RM, Karssemeijer N. Quantification of masking risk in screening mammography with volumetric breast density maps. *Breast Cancer Research & Treatment* 2017; **162**(3): 541-8.
62. Kerlikowske K, Zhu W, Tosteson AN, et al. Identifying women with dense breasts at high risk for interval cancer: a cohort study.[Summary for patients in Ann Intern Med. 2015 May 19;162(10). doi: 10.7326/P15-9018; PMID: 25984867]. *Annals of Internal Medicine* 2015; **162**(10): 673-81.
63. Smith V, Devane D, Begley C, Clarke M. Methodology in conducting a systematic review of systematic reviews of healthcare interventions. *BMC Medical Research Methodology* 2011; **11**(15).
64. Cummings SR, Tice JA, Bauer S, et al. Prevention of breast cancer in postmenopausal women: approaches to estimating and reducing risk. *Journal of the National Cancer Institute* 2009; **101**(6): 384-98.
65. Bae JM, Kim EH. Breast Density and Risk of Breast Cancer in Asian Women: A Meta-analysis of Observational Studies. *Journal of Preventive Medicine & Public Health / Yebang Uihakhoe Chi* 2016; **49**(6): 367-75.

66. Huo CW, Chew GL, Britt KL, et al. Mammographic density-a review on the current understanding of its association with breast cancer. *Breast Cancer Research & Treatment* 2014; **144**(3): 479-502.
67. Elias SG, Adams A, Wisner DJ, et al. Imaging features of HER2 overexpression in breast cancer: a systematic review and meta-analysis. *Cancer Epidemiology, Biomarkers & Prevention* 2014; **23**(8): 1464-83.
68. Antoni S, Sasco AJ, dos Santos Silva I, McCormack V. Is mammographic density differentially associated with breast cancer according to receptor status? A meta-analysis. *Breast Cancer Research & Treatment* 2013; **137**(2): 337-47.
69. McCormack VA, dos Santos Silva I. Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis. *Cancer Epidemiology, Biomarkers & Prevention* 2006; **15**(6): 1159-69.
70. Chang JM, Koo HR, Moon WK. Radiologist-performed hand-held ultrasound screening at average risk of breast cancer: results from a single health screening center. *Acta Radiologica* 2015; **56**(6): 652-8.
71. Destounis S, Arieno A, Morgan R. Initial experience with the New York State breast density inform law at a community-based breast center. *Journal of Ultrasound in Medicine* 2015; **34**(6): 993-1000.
72. Hwang JY, Han BK, Ko EY, Shin JH, Hahn SY, Nam MY. Screening Ultrasound in Women with Negative Mammography: Outcome Analysis. *Yonsei Medical Journal* 2015; **56**(5): 1352-8.
73. Kim SY, Kim MJ, Moon HJ, Yoon JH, Kim EK. Application of the downgrade criteria to supplemental screening ultrasound for women with negative mammography but dense breasts. *Medicine* 2016; **95**(44): e5279.
74. Klevos GA, Collado-Mesa F, Net JM, Yepes MM. Utility of supplemental screening with breast ultrasound in asymptomatic women with dense breast tissue who are not at high risk for breast cancer. *Indian Journal of Radiology & Imaging* 2017; **27**(1): 52-8.
75. Moon HJ, Jung I, Park SJ, Kim MJ, Youk JH, Kim EK. Comparison of Cancer Yields and Diagnostic Performance of Screening Mammography vs. Supplemental Screening Ultrasound in 4394 Women with Average Risk for Breast Cancer. *Ultraschall in der Medizin* 2015; **36**(3): 255-63.
76. Tagliafico AS, Calabrese M, Mariscotti G, et al. Adjunct Screening With Tomosynthesis or Ultrasound in Women With Mammography-Negative Dense Breasts: Interim Report of a Prospective Comparative Trial. *Journal of Clinical Oncology* 2016; **09**: 09.
77. Weigert J, Steenbergen S. The connecticut experiments second year: ultrasound in the screening of women with dense breasts. *Breast Journal* 2015; **21**(2): 175-80.
78. Wilczek B, Wilczek HE, Rasouliyan L, Leifland K. Adding 3D automated breast ultrasound to mammography screening in women with heterogeneously and extremely dense breasts: Report from a hospital-based, high-volume, single-center breast cancer screening program. *European Journal of Radiology* 2016; **85**(9): 1554-63.
79. Brem R, Tabár L, Duffy S, et al. Assessing improvement in detection of breast cancer with three-dimensional automated breast US in women with dense breast tissue: the SomoInsight Study. *Radiology* 2015; **274**(3): 663-73.
80. Giuliano V, Giuliano C. Improved breast cancer detection in asymptomatic women using 3D-automated breast ultrasound in mammographically dense breasts. *Clinical Imaging* 2013; **37**(3): 480-6.
81. Weigert J, Steenbergen S. The connecticut experiment: the role of ultrasound in the screening of women with dense breasts. *Breast Journal* 2012; **18**(6): 517-22.

82. Weigert JM. The Connecticut Experiment; The Third Installment: 4 Years of Screening Women with Dense Breasts with Bilateral Ultrasound. *Breast Journal* 2017; **23**(1): 34-9.
83. Destounis S, Arieno A, Morgan R. New York State Breast Density Mandate: Follow-up Data With Screening Sonography. *Journal of Ultrasound in Medicine* 2017; **28**: 28.
84. Gray E, Donten A, Karssemeijer N, et al. Evaluation of a Stratified National Breast Screening Program in the United Kingdom: An Early Model-Based Cost-Effectiveness Analysis. *Value in Health* 2017.
85. Sprague BL, Stout NK, Schechter C, et al. Benefits, harms, and cost-effectiveness of supplemental ultrasonography screening for women with dense breasts. *Annals of Internal Medicine* 2015; **162**(3): 157-66.
86. American College of Radiology. ACR Statement on Reporting Breast Density in Mammography Reports and Patient Summaries. 2017. https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Reporting-Breast-Density.
87. Duffy SW, Morrish OWE, Allgood PC, et al. Mammographic density and breast cancer risk in breast screening assessment cases and women with a family history of breast cancer. *Eur J Cancer* 2018; **88**: 48-56.
88. Lee CI, Chen LE, Elmore JG. Risk-based Breast Cancer Screening: Implications of Breast Density. *Medical Clinics of North America* 2017; **101**(4): 725-41.
89. Geisel J, Raghu M, Hooley R. The Role of Ultrasound in Breast Cancer Screening: The Case for and Against Ultrasound. *Seminars in Ultrasound, CT & MR* 2018; **39**(1): 25-34.

Appendix 1 Search strategy

Breast ultrasound searches for Q1, Q2 and Q3 in Medline and Embase were run up to July 10 2017.

Question 1: What are the reliability and validity of available methods to measure mammographic breast density?

Medline/Embase from 2000

1. (breast* adj2 dens*).tw.
2. (mammogra* adj2 dens*).tw.
3. Breast Density/
4. volumetric breast composition.mp.
5. 1 or 2 or 3 or 4
6. (Volpara* or cumulus or imageJ* or quantra or Single energy x-ray absorptiometry or DM-Density or M-Vu Breast).tw.
7. Ultrasonography, Mammary/
8. (ultrasound or ultrasonograph* or ultrasonic* or sonograph*).tw.
9. exp Mammography/
10. (BIRADS or BI-RADS).tw.
11. mammograph*.tw.
12. 6 or 7 or 8 or 9 or 10 or 11
13. 5 and 12
14. exp "Reproducibility of Results"/
15. exp observer variation/
16. (reliability or reliable or valid* or evaluat* or measure* or variability or variation or intra-rater or consisten* or performance or concordan* or discordan* or agreement or correlat* or reproducib*).tw.
17. 14 or 15 or 16
18. 13 and 17
19. limit 18 to english language

Cochrane Central Register of Controlled Trials : Issue 9, September 2017

Search strategy: 'mammogra* AND screen* AND (breast density OR dense breast* OR parenchym*) in Title, Abstract, Keywords, Publication Year from 2015 to 2017 in Trials.

Question 2. Is mammographic breast density a risk factor for cancers being missed during screening (false negatives/interval cancers)?

Medline/Embase from 2000

1. (breast* adj2 dens*).tw.
2. (mammogra* adj2 dens*).tw.

3. Breast Density/

4. volumetric breast composition.mp. [mp=title, abstract, heading word, drug trade name, original title, device manufacturer, drug manufacturer, device trade name, keyword, floating subheading word]

5. 1 or 2 or 3 or 4

6. exp Breast Neoplasms/cl, di, dg [Classification, Diagnosis, Diagnostic Imaging]

7. (breast adj2 (cancer or carcinoma or DCIS or malignan*)).tw.

8. "Early Detection of Cancer"/

9. 6 or 7 or 8

10. 5 and 9

11. risk.mp. or Risk/

12. (associated or association or relationship or odds ratio).tw.

13. 11 or 12

14. 10 and 13

15. limit 14 to english language

16. conference.pt.

17. 15 not 16

Question 3. What is the test accuracy of ultrasound in comparison to mammography in women with dense breasts?

Medline

1. (breast* adj2 dens*).tw.

2. (mammogra* adj2 dens*).tw.

3. Breast Density/

4. volumetric breast composition.mp.

5. 1 or 2 or 3 or 4

6. (Volpara* or cumulus or imageJ* or quantra or Single energy x-ray absorptiometry or DM-Density or M-Vu Breast).tw.

7. Ultrasonography, Mammary/

8. (ultrasound or ultrasonograph* or ultrasonic* or sonograph*).tw.

9. exp Mammography/

10. (BIRADS or BI-RADS).tw.

11. mammograph*.tw.

12. 6 or 7 or 8 or 9 or 10 or 11

13. 5 and 12

14. (detect* or specific* or sensitive* or accura* or predict* or false-positive or false-negative or true-negative or true-positive or AUC or ROC or odds ratio or performance).tw.

15. exp "Sensitivity and Specificity"/

16. 14 or 15

17. 13 and 16

18. limit 17 to english language

19. 6 or 7 or 8

20. 9 or 10 or 11

21. 5 and 19 and 20

22. limit 21 to english language

23. 18 or 22

Embase

1. (breast* adj2 dens*).tw.

2. (mammogra* adj2 dens*).tw.

3. Breast Density/

4. volumetric breast composition.mp.

5. 1 or 2 or 3 or 4

6. (Volpara* or cumulus or imageJ* or quantra or Single energy x-ray absorptiometry or DM-Density or M-Vu Breast).tw.

7. Ultrasonography, Mammary/

8. (ultrasound or ultrasonograph* or ultrasonic* or sonograph*).tw.

9. exp Mammography/

10. (BIRADS or BI-RADS).tw.

11. mammograph*.tw.

12. 6 or 7 or 8 or 9 or 10 or 11

13. 5 and 12

14. (detect* or specific* or sensitive* or accura* or predict* or false-positive or false-negative or true-negative or true-postive or AUC or ROC or odds ratio or performance).tw.

15. exp "Sensitivity and Specificity"/

16. 14 or 15

17. 13 and 16

18. limit 17 to english language

19. 6 or 7 or 8

20. 9 or 10 or 11

21. 5 and 19 and 20

22. limit 21 to english language

23. 18 or 22

24. conference.pt.

25. 23 not 24

Question 4. For women attending breast screening in the UK, what are the cost-consequences of adding mammographic density measurements, and then ultrasound for those found to have high mammographic breast density?

Medline

Searched Ovid MEDLINE(R) 1946 to January Week 2 2018, Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations January 22, 2018, Ovid MEDLINE(R) Epub Ahead of Print January 22, 2018

1. (breast* adj2 dens*).tw.

2. (mammogra* adj2 dens*).tw.

3. Breast Density/

4. volumetric breast composition.mp.

5. 1 or 2 or 3 or 4

6. (Volpara* or cumulus or imageJ* or quantra or Single energy x-ray absorptiometry or DM-Density or M-Vu Breast).tw.

7. Ultrasonography, Mammary/

8. (ultrasound or ultrasonograph* or ultrasonic* or sonograph*).tw.

9. exp Mammography/

10. (BIRADS or BI-RADS).tw.

11. mammograph*.tw.

12. 6 or 7 or 8 or 9 or 10 or 11

13. 5 and 12

14. exp Economics/

15. exp "Costs and Cost Analysis"/

16. exp Quality-Adjusted Life Years/

17. (pharmacoeconomic* or pharmaco-economic* or economic* or cost*).tw.

18. (qaly* or ICER* or utilit* or EQ5D* or EQ-5D* or euroqol* or euro-qol* or short form or SF-36 or SF36 or SF-6D or SF6D or SF-12 or SF12 or HUI).tw.

19. (decision adj2 model).tw.

20. ((resource* adj2 utili$ation) or 'resource use').tw.

21. (utilit* adj2 (value* or index* or health or measure* or estimate*)).tw.

22. 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21

23. 13 and 22

24. limit 23 to english language

25. limit 24 to yr="2005 -Current"

135 downloaded

Embase

Ovid Embase 1980 to 2018 Week 04

1. (breast* adj2 dens*).tw.

2. (mammogra* adj2 dens*).tw.

3. Breast Density/

4. volumetric breast composition.mp.

5. 1 or 2 or 3 or 4

6. (Volpara* or cumulus or imageJ* or quantra or Single energy x-ray absorptiometry or DM-Density or M-Vu Breast).tw.

7. Ultrasonography, Mammary/

8. (ultrasound or ultrasonograph* or ultrasonic* or sonograph*).tw.

9. exp Mammography/

10. (BIRADS or BI-RADS).tw.

11. mammograph*.tw.

12. 6 or 7 or 8 or 9 or 10 or 11

13. 5 and 12

14. exp Economics/

15. exp "Costs and Cost Analysis"/

16. exp Quality-Adjusted Life Years/

17. (pharmacoeconomic* or pharmaco-economic* or economic* or cost*).tw.

18. (qaly* or ICER* or utilit* or EQ5D* or EQ-5D* or euroqol* or euro-qol* or short form or SF-36 or SF36 or SF-6D or SF6D or SF-12 or SF12 or HUI).tw.

19. (decision adj2 model).tw.

20. ((resource* adj2 utili$ation) or 'resource use').tw.

21. (utilit* adj2 (value* or index* or health or measure* or estimate*)).tw.

22. 14 or 15 or 16 or 17 or 18 or 19 or 20 or 21

23. 13 and 22

24. limit 23 to english language

25. limit 24 to yr="2005 -Current"

26. conference abstract.pt.

27. 25 not 26

165 downloaded

Web of Science Core Collection

125

Searched: TOPIC: (breast* NEAR/3 dens*) AND TOPIC: (ultrasound or ultrasonograph* or ultrasonic* or sonograph* or supplemental) AND TOPIC: (cost* or economic* or QALY*) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article)

Timespan: 2005-2018. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI.

51 downloaded

Cochrane Library (23/01/2018) NHS Economic Evaluation Database and Health Technology Assessment Database :

Searched: 'breast* near/3 dens* in Title, Abstract, Keywords and cost* or economic* or QALY* in Title, Abstract, Keywords

8 records downloaded

Cost-effectiveness Analysis (CEA) Registry

4 records found

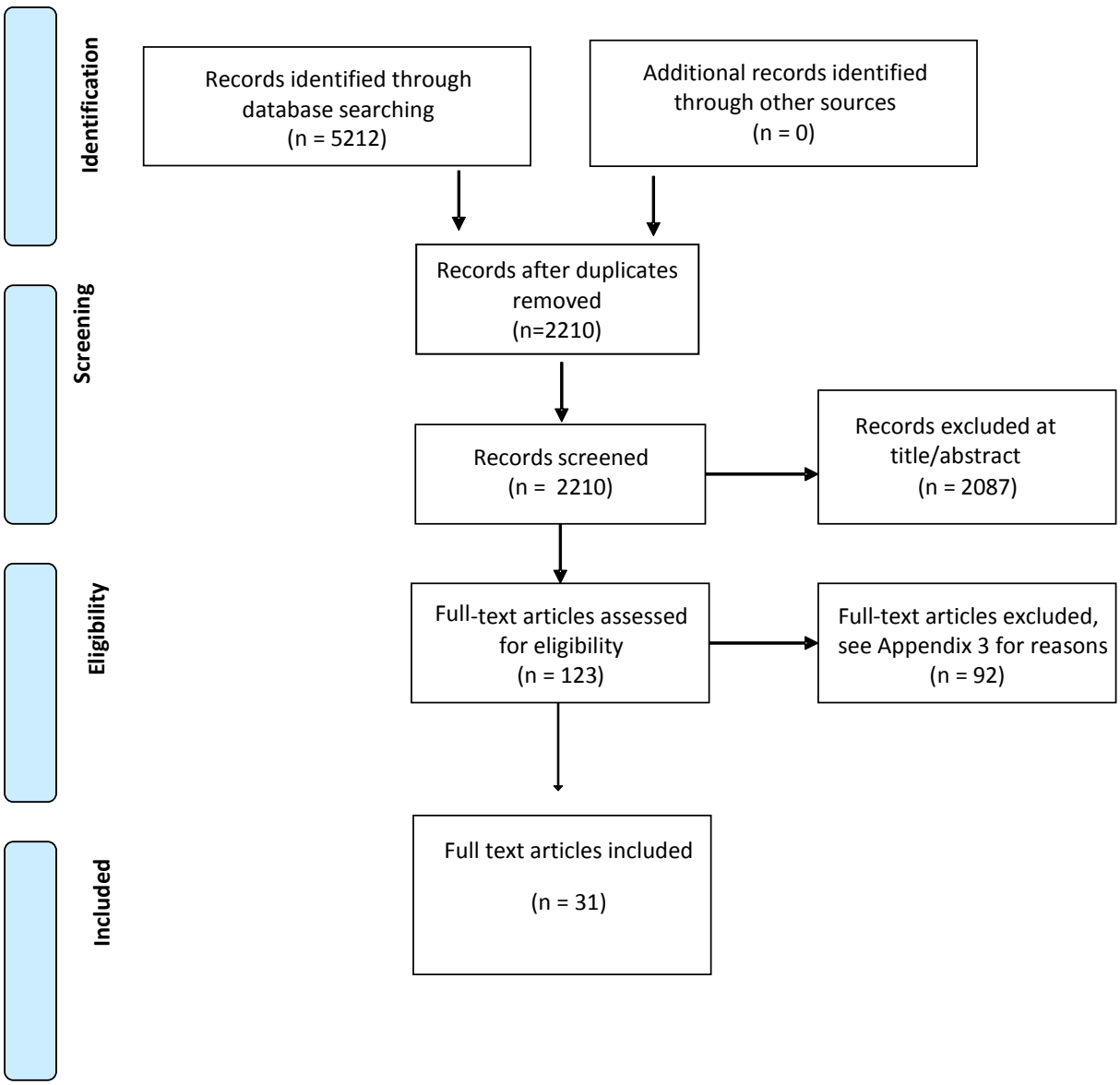
Endnote: total of 359 records before deduplication; After deduplication = 201

Searches were supplemented with weekly database auto-alerts and update searches; papers identified by experts; and examining reference lists of identified papers.

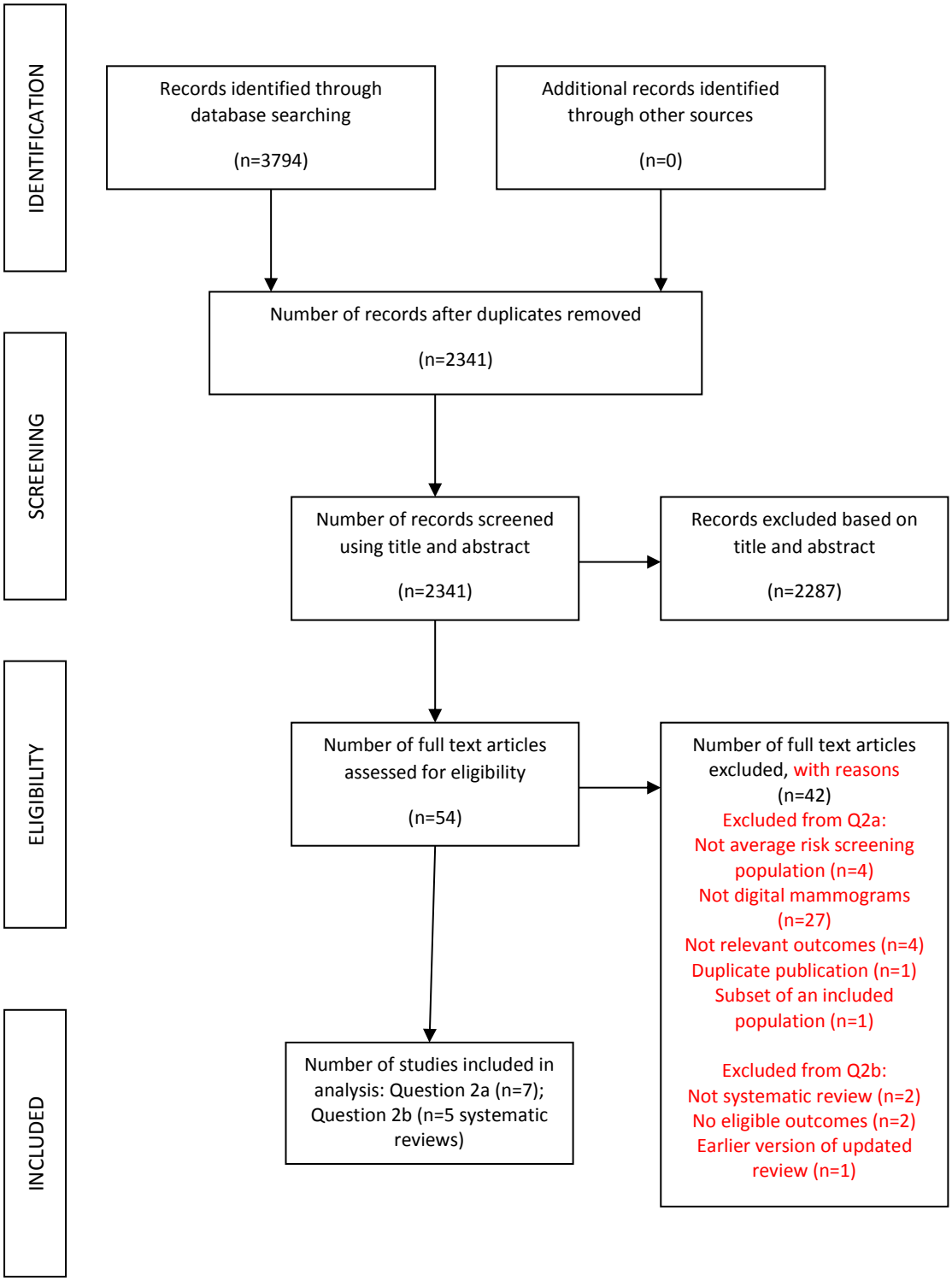
Appendix 2 PRISMA record selection

Question 1

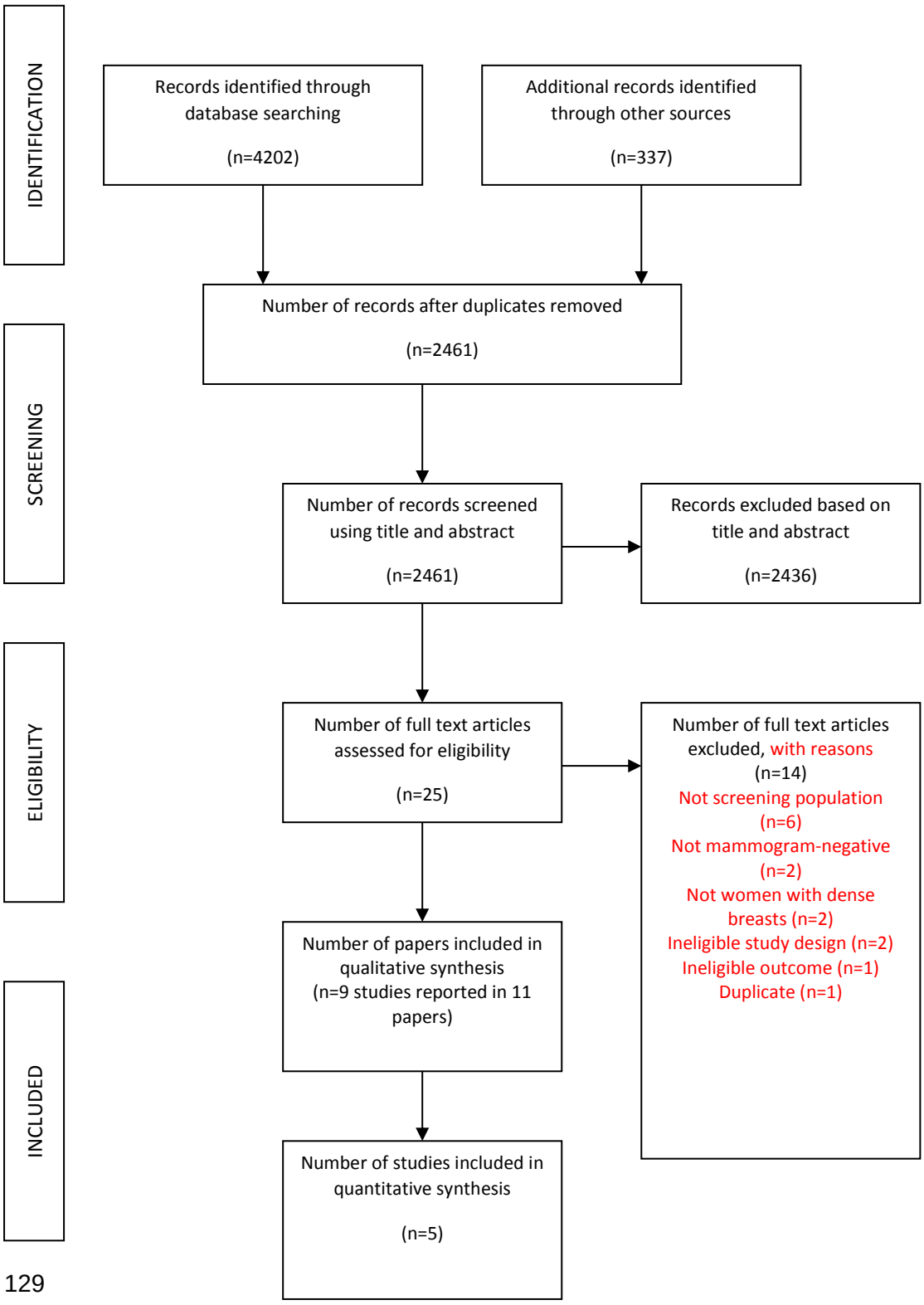
PRISMA flow chart for question 1



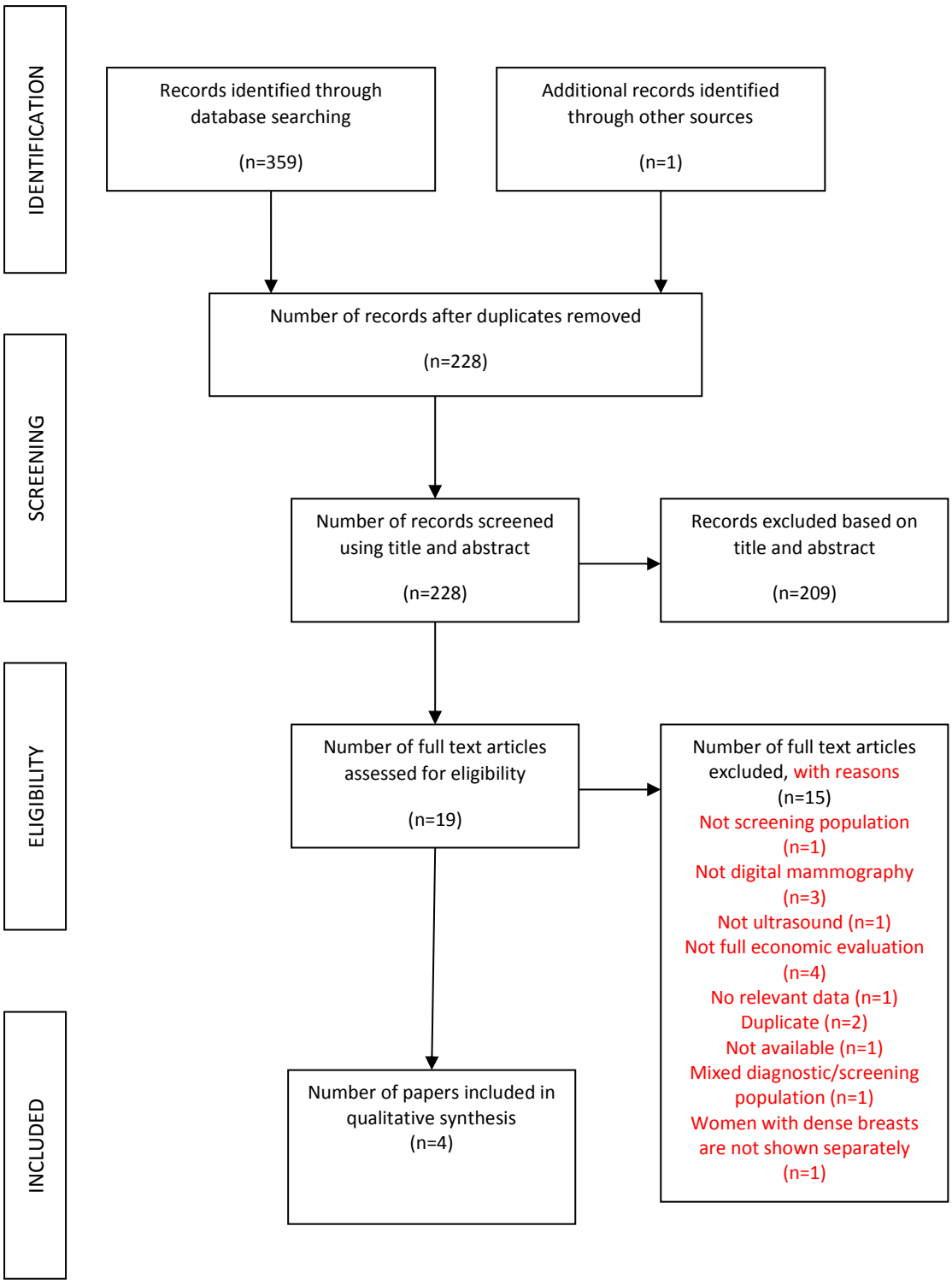
Question 2


IDENTIFICATION
SCREENING
ELIGIBILITY
INCLUDED
Records identified through database searching (n=3794)
Additional records identified through other sources (n=0)
Number of records after duplicates removed (n=2341)
Number of records screened using title and abstract (n=2341)
Records excluded based on title and abstract (n=2287)
Number of full text articles assessed for eligibility (n=54)
Number of full text articles excluded, with reasons (n=42)
Excluded from Q2a:
Not average risk screening population (n=4)
Not digital mammograms (n=27)
Not relevant outcomes (n=4)
Duplicate publication (n=1)
Subset of an included population (n=1)
Excluded from Q2b:
Not systematic review (n=2)
No eligible outcomes (n=2)
Earlier version of updated review (n=1)
Number of studies included in analysis: Question 2a (n=7); Question 2b (n=5 systematic reviews)

Question 3


IDENTIFICATION
SCREENING
ELIGIBILITY
INCLUDED
Records identified through database searching (n=4202)
Additional records identified through other sources (n=337)
Number of records after duplicates removed (n=2461)
Number of records screened using title and abstract (n=2461)
Records excluded based on title and abstract (n=2436)
Number of full text articles assessed for eligibility (n=25)
Number of full text articles excluded, with reasons (n=14)
Not screening population (n=6)
Not mammogram-negative (n=2)
Not women with dense breasts (n=2)
Ineligible study design (n=2)
Ineligible outcome (n=1)
Duplicate (n=1)
Number of papers included in qualitative synthesis (n=9 studies reported in 11 papers)
Number of studies included in quantitative synthesis (n=5)

Question 4


IDENTIFICATION
SCREENING
ELIGIBILITY
INCLUDED
Records identified through database searching (n=359)
Additional records identified through other sources (n=1)
Number of records after duplicates removed (n=228)
Number of records screened using title and abstract (n=228)
Records excluded based on title and abstract (n=209)
Number of full text articles assessed for eligibility (n=19)
Number of full text articles excluded, with reasons (n=15)
Not screening population (n=1)
Not digital mammography (n=3)
Not ultrasound (n=1)
Not full economic evaluation (n=4)
No relevant data (n=1)
Duplicate (n=2)
Not available (n=1)
Mixed diagnostic/screening population (n=1)
Women with dense breasts are not shown separately (n=1)
Number of papers included in qualitative synthesis (n=4)

Appendix 3 Excluded studies

Question 1

Paper	Reason for exclusion
Abdolell, M., et al. (2016). "Consistency of visual assessments of mammographic breast density from vendor-specific "for presentation" images." Journal of Medical Imaging 3(1): 011004.	Diagnostic mammograms
Alipour, S., et al. (2013). "Imperfect correlation of mammographic and clinical breast tissue density." Asian Pacific Journal of Cancer Prevention: Apjcp 14(6): 3685-3688.	Ineligible comparator
Benichou, J., et al. (2003). "Secular stability and reliability of measurements of the percentage of dense tissue on mammograms." Cancer Detection & Prevention 27(4): 266-274.	Film mammography
Berg, W. A., et al. (2000). "Breast Imaging Reporting and Data System: inter- and intraobserver variability in feature analysis and final assessment." AJR. American Journal of Roentgenology 174(6): 1769-1777.	Film mammography
Bernardi, D., et al. (2012). "Interobserver agreement in breast radiological density attribution according to BI-RADS quantitative classification." Radiologia Medica 117(4): 519-528.	Film mammography
Brandt, K. R., et al. (2016). "Comparison of Clinical and Automated Breast Density Measurements: Implications for Risk Prediction and Supplemental Screening." Radiology 279(3): 710-719.	Multiple cohorts
Brentnall, A. R., et al. (2015). "Mammographic density adds accuracy to both the Tyrer-Cuzick and Gail breast cancer risk models in a prospective UK screening cohort." Breast Cancer Research 17(1): 147.	Mixed film/digital mammography
Burton, A., et al. (2016). "Mammographic density assessed on paired raw and processed digital images and on paired screen-film and digital images across three mammography systems." Breast Cancer Research 18(1): 130.	Multiple cohorts
Busana, M. C., et al. (2016). "Assessing within-woman changes in mammographic density: a comparison of fully versus semi-automated area-based approaches." Cancer Causes & Control 27(4): 481-491.	Film mammography
Castillo-Garcia, M., et al. (2017). "Automated Breast Density Computation in Digital Mammography and Digital Breast Tomosynthesis: Influence on Mean Glandular Dose and BIRADS Density Categorization." Academic Radiology 24(7): 802-810.	Mixed opportunistic screening/diagnostic population
Chang, R. F., et al. (2006). "Breast density analysis in 3-D whole breast ultrasound images." Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine & Biology Society 1: 2795-2798.	<100 women
Chang, R. F., et al. (2006). "Three comparative approaches for breast density estimation in digital and screen film mammograms." Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine & Biology Society 1: 4853-4856.	<100 women
Chang, Y. H., et al. (2002). "Computerized assessment of tissue composition on digitized mammograms." Academic Radiology 9(8): 899-905.	Film mammography
Cheddad, A., et al. (2014). "Area and volumetric density estimation in processed full-field digital mammograms for risk assessment of breast cancer." PLoS ONE [Electronic Resource] 9(10): e110690.	Ineligible comparator
Ciatto, S., et al. (2012). "A first evaluation of breast radiological density assessment by QUANTRA software as compared to visual classification." Breast 21(4): 503-506.	Symptomatic/spontaneous screening, and breast surgery population

Ciatto, S., et al. (2005). "Categorizing breast mammographic density: intra- and interobserver reproducibility of BI-RADS density categories." Breast 14(4): 269-275.	Film mammography
Couwenberg, A. M., et al (2014). “Assessment of a fully automated, high-throughput mammographic density measurement tool for use with processed digital mammograms.” Cancer Causes & Control 25(8): 1037-43.	Correlates ImageJ with Cumulus (which was used to train the ImageJ) in the training set; validation set not screening population
Damases, C. N., et al. (2015). "Mammographic density measurements are not affected by mammography system." Journal of Medical Imaging 2(1): 015501.	<100 women
Damases, C. N., et al (2017). “Intercountry analysis of breast density classification using visual grading.” Br J Radiol 90(1076): 20170064	<100 women
Ekpo, E. U., et al. (2016). "Assessment of Interradiologist Agreement Regarding Mammographic Breast Density Classification Using the Fifth Edition of the BI-RADS Atlas." AJR. American Journal of Roentgenology 206(5): 1119-1123.	Mixed screening/diagnostic population
Ekpo, E. U., et al. (2017). "A self-directed learning intervention for radiographers rating mammographic breast density." Radiography. 10.	<100 women
Engelken 2014. Volumetric breast composition analysis: reproducibility of breast percent density and fibroglandular tissue volume measurements in serial mammograms. Acta Radiologica 55(1): 32-8	No eligible data
Gao, J., et al. (2008). "Reproducibility of visual assessment on mammographic density." Breast Cancer Research & Treatment 108(1): 121-127.	Participants at high risk of cancer
Gard, C. C., et al. (2015). "Misclassification of Breast Imaging Reporting and Data System (BI-RADS) Mammographic Density and Implications for Breast Density Reporting Legislation." Breast Journal 21(5): 481-489.	Film mammography
Glide-Hurst, C. K., et al. (2007). "A new method for quantitative analysis of mammographic density." Medical Physics 34(11): 4491-4498.	Film mammography
Gram, I. T., et al. (2005). "Percentage density, Wolfe's and Tabar's mammographic patterns: agreement and association with risk factors for breast cancer." Breast Cancer Research 7(5): R854-861.	Film mammography
Heine, J. J., et al. (2011). "Calibrated measures for breast density estimation." Academic Radiology 18(5): 547-555.	Ineligible comparator
Heine, J. J., et al. (2011). "A quantitative description of the percentage of breast density measurement using full-field digital mammography." Academic Radiology 18(5): 556-564.	Ineligible comparator
Hersh, M. A. (2004). "Imaging the dense breast." Applied Radiology 33(1): 22-26.	Excluded study design: summary of density
Highnam, R., et al. (2007). "Comparing measurements of breast density." Physics in Medicine & Biology 52(19): 5881-5895.	Ineligible interventions
Highnam, R., et al. (2006). "Breast composition measurements using retrospective standard mammogram form (SMF)." Physics in Medicine & Biology 51(11): 2695-2713.	Ineligible interventions
Hodge, R., et al. (2014). "Comparison of Danish dichotomous and BI-RADS classifications of mammographic density." Acta Radiologica Short Reports 3(5): 2047981614536558.	Film mammography
Iatrakis, G., et al. (2010). "Preliminary results of objective assessment of mammographic percent density." Clinical & Experimental Obstetrics & Gynecology 37(1): 24-25.	<100 women
Iatrakis, G., et al. (2011). "Quantitative assessment of breast mammographic density with a new objective method." Journal of Medicine & Life 4(3): 310-313.	<100 women
Jamal, N., et al. (2006). "Quantitative assessment of breast density from digitized mammograms into Tabar's patterns." Physics in Medicine & Biology 51(22): 5843-5857.	Diagnostic mammography

Jari, I., et al. (2014). "Computerized calculation of breast density: our experience from Arcadia Medical Imaging Center." Revista Medico-Chirurgicala a Societatii de Medici Si Naturalisti Din Iasi 118(4): 979-985.	Film mammography
Jeffreys, M., et al. (2006). "Initial experiences of using an automated volumetric measure of breast density: the standard mammogram form." British Journal of Radiology 79(941): 378-382.	Ineligible interventions
Kallenberg, M. G., et al. (2011). "Automatic breast density segmentation: an integration of different approaches." Physics in Medicine & Biology 56(9): 2715-2729.	Film mammography
Kataoka, M., et al. (2008). "Mammographic density using two computer-based methods in an isoflavone trial." Maturitas 59(4): 350-357.	Film mammography
Keller, B. M., et al. (2015). "Preliminary evaluation of the publicly available Laboratory for Breast Radiodensity Assessment (LIBRA) software tool: comparison of fully automated area and volumetric density measures in a case-control study with digital mammography." Breast Cancer Research 17: 117.	Mixed population (screening and diagnostic)
Kim, W. H., et al. (2013). "Variability of breast density assessment in short-term reimaging with digital mammography." European Journal of Radiology 82(10): 1724-1730.	Not screening population
Ko, S. Y., et al. (2014). "Mammographic density estimation with automated volumetric breast density measurement." Korean Journal of Radiology 15(3): 313-321.	Mixed population (screening and diagnostic)
Kotsuma, Y., et al. (2008). "Quantitative assessment of mammographic density and breast cancer risk for Japanese women." Breast 17(1): 27-35.	Film mammography
Lee, H. N., et al. (2015). "Comparison of mammographic density estimation by Volpara software with radiologists' visual assessment: analysis of clinical-radiologic factors affecting discrepancy between them." Acta Radiologica 56(9): 1061-1068.	Mixed population (screening and diagnostic)
Li, J., et al. (2012). "High-throughput mammographic-density measurement: a tool for risk prediction of breast cancer." Breast Cancer Research 14(4): R114.	Film mammography
Lokate, M., et al. (2010). "Volumetric breast density from full-field digital mammograms and its association with breast cancer risk factors: A comparison with a threshold method." Cancer Epidemiology Biomarkers and Prevention 19(12): 3096-3105.	Ineligible comparator
Lu, L. J., et al. (2007). "Computing mammographic density from a multiple regression model constructed with image-acquisition parameters from a full-field digital mammographic unit." Physics in Medicine & Biology 52(16): 4905-4921.	Ineligible comparator
Machida, Y., et al. (2016). "Automated volumetric breast density estimation out of digital breast tomosynthesis data: feasibility study of a new software version." Springerplus 5(1): 780.	Ineligible comparator
Marias, K., et al. (2005). "Automatic labelling and BI-RADS characterisation of mammogram densities." Conference Proceedings: ... Annual International Conference of the IEEE Engineering in Medicine & Biology Society 6: 6394-6398.	Excluded study design: description of method of automated characterisation of density
Maskarinec, G., et al. (2011). "Comparison of breast density measured by dual energy X-ray absorptiometry with mammographic density among adult women in Hawaii." Cancer Epidemiology 35(2): 188-193.	Film mammography
Masroor, I., et al. (2016). "To asses inter- and intra-observer variability for breast density and BIRADS assessment categories in mammographic reporting." JPMA - Journal of the Pakistan Medical Association 66(2): 194-197.	Mixed population (screening and diagnostic)
McCormack, V. A., et al. (2007). "Comparison of a new and existing method of mammographic density measurement: intramethod	Film mammography

reliability and associations with known risk factors." Cancer Epidemiology, Biomarkers & Prevention 16(6): 1148-1154.	
Meggiorini, M. L., et al. (2016). "Mammographic breast density in infertile and parous women." BMC Women's Health 16: 8.	High risk population
Moradi, M., et al. (2013). "Performance of double reading mammography in an Iranian population and its effect on patient outcome." Iranian Journal of Radiology 10(2): 51-55.	Film mammography
Morrish, O. W., et al. (2015). "Mammographic breast density: comparison of methods for quantitative evaluation." Radiology 275(2): 356-365.	High risk population
Ng, K. H., et al. (2012). "Standardisation of clinical breast-density measurement." Lancet Oncology 13(4): 334-336.	Excluded study design: comment
Nicholson, B. T., et al. (2006). "Accuracy of assigned BI-RADS breast density category definitions." Academic Radiology 13(9): 1143-1149.	Film mammography
Nithya, R. and B. Santhi (2017). "Computer-aided diagnosis system for mammogram density measure and classification." Biomedical Research (India) 28(6): 2427-2431.	Not reliability/concordance
Oliver, A., et al. (2008). "A novel breast tissue density classification methodology." IEEE Transactions on Information Technology in Biomedicine 12(1): 55-65.	Film mammography
Oliver, A., et al. (2006). "A comparison of breast tissue classification techniques." Medical Image Computing & Computer-Assisted Intervention: MICCAI 9(Pt 2): 872-879.	Ineligible comparator
Oliver, A., et al. (2010). "Influence of using manual or automatic breast density information in a mass detection CAD system." Academic Radiology 17(7): 877-883.	<100 women
Pahwa, S., et al. (2015). "Evaluation of breast parenchymal density with QUANTRA software." Indian Journal of Radiology & Imaging 25(4): 391-396.	Not a screening population (no screening programme in place, mixed self-referral/diagnostic)
Pawluczyk, O., et al. (2003). "A volumetric method for estimation of breast density on digitized screen-film mammograms." Medical Physics 30(3): 352-364.	Film mammography
Perez-Gomez, B., et al. (2011). “Women's features and inter-/intra-rater agreement on mammographic density assessment in full-field digital mammograms (DDM-SPAIN).” Breast Cancer Research and Treatment: 1-9	No appropriate interventions
Prevrhal, S., et al. (2002). "Accuracy of mammographic breast density analysis: results of formal operator training." Cancer Epidemiology, Biomarkers & Prevention 11(11): 1389-1393.	Film mammography
Redondo, A., et al. (2012). "Inter- and intraradiologist variability in the BI-RADS assessment and breast density categories for screening mammograms." British Journal of Radiology 85(1019): 1465-1470.	Film mammography
Regini, E., et al. (2014). "Radiological assessment of breast density by visual classification (BI-RADS) compared to automated volumetric digital software (Quantra): implications for clinical practice." Radiologia Medica 119(10): 741-749.	Mixed self-referred to screening/diagnostic population
Sacchetto, D., et al. (2015). "Mammographic density: Comparison of visual assessment with fully automatic calculation on a multivendor dataset." Journal of Nanoparticle Research 17(12): 175-183.	Duplicate
Sacchetto, D., et al. (2016). "Mammographic density: Comparison of visual assessment with fully automatic calculation on a multivendor dataset." European Radiology 26(1): 175-183.	Mixed population (screening and diagnostic)
Sawada, T., et al. (2017). "Digital volumetric measurement of mammographic density and the risk of overlooking cancer in Japanese women." Breast Cancer: 1-6.	Not reliability/concordance

Schmachtenberg, C., et al. (2015). "Intraindividual comparison of two methods of volumetric breast composition assessment." Academic Radiology 22(4): 447-452.	Not screening population
Shepherd, J. A., et al. (2005). "Novel use of single X-ray absorptiometry for measuring breast density." Technology in Cancer Research & Treatment 4(2): 173-182.	Film mammography
Shepherd, J. A., et al. (2011). "Volume of mammographic density and risk of breast cancer." Cancer Epidemiology, Biomarkers & Prevention 20(7): 1473-1482.	Not reliability/concordance
Singh, J. M., et al. (2013). "Volumetric breast density assessment: reproducibility in serial examinations and comparison with visual assessment." Rofo: Fortschritte auf dem Gebiete der Rontgenstrahlen und der Nuklearmedizin 185(9): 844-848.	Not screening population (surveillance after breast surgery or diagnostic)
Soares, D., et al. (2002). "Age as a predictive factor of mammographic breast density in Jamaican women." Clinical Radiology 57(6): 472-476.	Mixed population (screening and diagnostic)
Sohn, G., et al. (2014). "Reliability of the percent density in digital mammography with a semi-automated thresholding method." Journal of Breast Cancer 17(2): 174-179.	Not screening population
Sperrin, M., et al. (2013). "Correcting for rater bias in scores on a continuous scale, with application to breast density." Statistics in Medicine 32(26): 4666-4678.	Film mammography
Sprague, B. L., et al. (2016). "Variation in Assessments of Breast Density on Mammograms in Clinical Practice." Annals of Internal Medicine 165 (7) (no pagination)(I-28).	Summary of a study for patients
Stone, J., et al. (2010). "Predicting breast cancer risk using mammographic density measurements from both mammogram sides and views." Breast Cancer Research & Treatment 124(2): 551-554.	Not reliability/concordance
Tagliafico, A., et al. (2009). "Mammographic density estimation: comparison among BI-RADS categories, a semi-automated software and a fully automated one." Breast 18(1): 35-40.	Film mammography
Tagliafico, A. S., et al. (2013). "Estimation of percentage breast tissue density: comparison between digital mammography (2D full field digital mammography) and digital breast tomosynthesis according to different BI-RADS categories." British Journal of Radiology 86(1031): 20130255.	Diagnostic population
Tomas, I., et al. (2013). "Computer-aided evaluation of radiologist's reproducibility and subjectivity in mammographic density assessment." Collegium Antropologicum 37(4): 1121-1126.	Film mammography
Trocchi, P., et al. (2012). "Mammographic density and inter-observer variability of pathologic evaluation of core biopsies among women with mammographic abnormalities." BMC Cancer 12: 554.	Not screening population
Vachon, C. M., et al. (2013). "Comparison of percent density from raw and processed full-field digital mammography data." Breast Cancer Research 15(1): R1.	Mixed population (screening and diagnostic)
Winkel, R. R., et al. (2015). "Inter-observer agreement according to three methods of evaluating mammographic density and parenchymal pattern in a case control study: impact on relative risk of breast cancer." BMC Cancer 15: 274.	Film mammography
Woolcott, C. G., et al. (2014). "Methods for assessing and representing mammographic density: an analysis of 4 case-control studies." American Journal of Epidemiology 179(2): 236-244.	Film mammography
Yan, S., et al. (2017). "Applying a new bilateral mammographic density segmentation method to improve accuracy of breast cancer risk prediction." International Journal of Computer Assisted Radiology and Surgery: 1-10.	Ineligible interventions
Youk, J. H., et al. (2016). "Automated Volumetric Breast Density Measurements in the Era of the BI-RADS Fifth Edition: A Comparison	Not screening population

With Visual Assessment." AJR. American Journal of Roentgenology 206(5): 1056-1062.	
Youk 2017. Comparison of Visual Assessment of Breast Density in BI-RADS 4th and 5th Editions With Automated Volumetric Measurement. American Journal of Roentgenology. 2017;209: 703-708.	Mixed population (screening and diagnostic)

Question 2

Study	Exclude reason
Bae 2014. Breast cancer detected with screening US: reasons for nondetection at mammography. Radiology 270(2): 369-77	Study showed that some cancers missed at mammography due to overlying dense tissue, but does not show the overall risk of missed cancer by density
Baglietto 2014. Associations of mammographic dense and nondense areas and body mass index with risk of breast cancer. American Journal of Epidemiology 179(4): 475-83	Film
Bare 2015. Mammographic and clinical characteristics of different phenotypes of screen-detected and interval breast cancers in a nationwide screening program. Breast Cancer Research & Treatment 154(2): 403-15	Film
Benichou 2003. Secular stability and reliability of measurements of the percentage of dense tissue on mammograms. Cancer Detection & Prevention 27(4): 266-74	Film screen or xeroradiogram
Blanch 2014. Impact of risk factors on different interval cancer subtypes in a population-based breast cancer screening programme. PLoS ONE. 9 (10) (no pagination): e110207	Mixed film/ digital
Chiarelli 2006. Influence of patterns of hormone replacement therapy use and mammographic density on breast cancer detection. Cancer Epidemiology, Biomarkers & Prevention 15(10): 1856-62	Film
Chiarelli 2015. Digital versus screen-film mammography: impact of mammographic density and hormone therapy on breast cancer detection. Breast Cancer Research & Treatment 2015; 154(2): 377-87.	No eligible outcomes
Chiu 2010. Effect of baseline breast density on breast cancer incidence, stage, mortality, and screening parameters: 25-year follow-up of a Swedish mammographic screening. Cancer Epidemiology, Biomarkers & Prevention. 19(5): 1219-28	Film
Choi 2016 Analysis of prior mammography with negative result in women with interval breast cancer. Breast Cancer 23(4): 583-9	Mixed film and digital
Ciatto 2004. Breast density as a determinant of interval cancer at mammographic screening. British Journal of Cancer 90(2): 393-6	Film
Collett 2005. A basal epithelial phenotype is more frequent in interval breast cancers compared with screen detected tumors. Cancer Epidemiology, Biomarkers & Prevention 14(5): 1108-12	Film
Domingo 2010. Phenotypic characterization and risk factors for interval breast cancers in a population-based breast cancer screening program in Barcelona, Spain. Cancer Causes & Control 21(8): 1155-64	Mixed film and digital
Domingo 2014. Tumor phenotype and breast density in distinct categories of interval cancer: results of population-based mammography screening in Spain. Breast Cancer Research 16(1): R3	Mixed film and digital

Elmore 2004. The association between obesity and screening mammography accuracy. Archives of Internal Medicine 164(10): 1140-7	Film
Henderson 2015. Performance of digital screening mammography among older women in the United States. Cancer 2015; 121 (9): 1379-86.	Subset of Nelson sample
Holm 2015. Risk factors and tumor characteristics of interval cancers by mammographic density. Journal of Clinical Oncology 33(9): 1030-1037	Film
Kavanagh 2008. Using mammographic density to improve breast cancer screening outcomes. Cancer Epidemiology, Biomarkers & Prevention 17(10): 2818-24	Film
Kim 2017. Analysis of Participant Factors That Affect the Diagnostic Performance of Screening Mammography: A Report of the Alliance for Breast Cancer Screening in Korea. Korean Journal of Radiology 18(4): 624-631	Not stated to be digital
Ko 2013. Comparison of new and established full-field digital mammography systems in diagnostic performance. Korean Journal of Radiology 14(2): 164-70	Mixed screening and high-risk women
Krishnan 2016. Mammographic density and risk of breast cancer by mode of detection and tumor size: a case-control study. Breast Cancer Research 18(1): 63	Film (same cohort as Baglietto)
Lowery 2011. Complementary approaches to assessing risk factors for interval breast cancer. Cancer Causes & Control 22(1): 23-31	Film
Malaj 2016. Synergy in combining findings from mammography and ultrasonography in detecting malignancy in women with higher density breasts and lesions over 2 cm in Albania. Wspolczesna Onkologia 2016; 20(6): 475-480	Not screening population
Mandelson 2000. Breast density as a predictor of mammographic detection: comparison of interval- and screen-detected cancers. Journal of the National Cancer Institute 92(13): 1081-7	Film
McDonald 2016. Performance of DWI as a Rapid Unenhanced Technique for Detecting Mammographically Occult Breast Cancer in Elevated-Risk Women With Dense Breasts. AJR. American Journal of Roentgenology 207(1): 205-16	Not density by interval cancer
Morimoto 2000. Breast cancer screening by mammography in women aged under 50 years in Japan. Anticancer Research 20(5C): 3689-94	Not stated to be digital (pre-March 1999)
Muttarak 2006. Breast carcinomas: why are they missed? Singapore Medical Journal 47(10): 851-7	Film
Nederend 2014. Impact of the transition from screen-film to digital screening mammography on interval cancer characteristics and treatment - a population based study from the Netherlands. European Journal of Cancer 2014; 50(1): 31-9	Does not report suitable outcomes
Nickson 2009. Tumour size at detection according to different measures of mammographic breast density. Journal of Medical Screening 16(3): 140-6	Film
Olsen 2009. Breast density and outcome of mammography screening: a cohort study. British Journal of Cancer 100(7): 1205-8	Film
Sanders 2016 (Screening subset). Impact of the New Jersey Breast Density Law on Imaging and Intervention Volumes and Breast Cancer	Mixed screening/high risk population

Diagnosis. Journal of the American College of Radiology 13(10): 1189-1194	
Sardanelli 2017. Position paper on screening for breast cancer by the European Society of Breast Imaging (EUSOBI) and 30 national breast radiology bodies from Austria, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Israel, Lithuania, Moldova, The Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Spain, Sweden, Switzerland and Turkey. European Radiology 27(7): 2737-2743	Question 2b but not a systematic review
Sawada 2017. Digital volumetric measurement of mammographic density and the risk of overlooking cancer in Japanese women. Breast Cancer 25: 25.	Mixed screening/ diagnostic population
Starikov 2016. 2D mammography, digital breast tomosynthesis, and ultrasound: which should be used for the different breast densities in breast cancer screening? Clinical Imaging 40(1): 68-71.	Question 2b but not a systematic review
van der Waal 2017. Breast cancer screening effect across breast density strata: A case-control study. International Journal of Cancer 140(1): 41-49	Film
Virnig 2009. Diagnosis and management of ductal carcinoma in situ (DCIS). Evidence Report/Technology Assessment 185: 1-549.	No eligible outcomes
Wanders 2017. The effect of volumetric breast density on the risk of screen-detected and interval breast cancers: a cohort study. Breast Cancer Research 19(1): 67	Duplicate (same cohort as Wanders 2017[7] with slightly fewer women)
Wang 2000. The evaluation of false negative mammography from malignant and benign breast lesions. Clinical Imaging 24(2): 96-103	Film
Wang 2001. Interval cancers in the Norwegian breast cancer screening program: frequency, characteristics and use of HRT. International Journal of Cancer 94(4): 594-8	Film
Wang 2013. Effects of age, breast density and volume on breast cancer diagnosis: a retrospective comparison of sensitivity of mammography and ultrasonography in China's rural areas. Asian Pacific Journal of Cancer Prevention: Apjcp 14(4): 2277-82	Not stated to be digital mammography
Weber 2016. Characteristics and prognosis of interval cancers after biennial screen-film or full-field digital screening mammography. Breast Cancer Research and Treatment 2016; 158(3): 471-483.	Not screening population – all had interval cancer
Weir R, et al. Risk factors for breast cancer in women. NZHTA Report 2007; 10(2).	No eligible outcomes (no unadjusted or only age-adjusted outcomes reported)
White 2004. Biennial versus annual mammography and the risk of late-stage breast cancer. Journal of the National Cancer Institute 96(24): 1832-9	Not stated to be digital

Question 3

Study	Reason for exclusion
Bowles 2016. The Use of Ultrasound in Breast Cancer Screening of Asymptomatic Women with Dense Breast Tissue: A Narrative Review. Journal of Medical Imaging and Radiation Sciences 47(3 Supplement): S21-S28	Systematic review

Brem 2015. Assessing improvement in detection of breast cancer with three-dimensional automated breast US in women with dense breast tissue: the SomoInsight Study. Radiology 274(3): 663-73	Duplicate (already included in USPTF review)
Dong, H., et al. Improved Performance of Adjunctive Ultrasonography After Mammography Screening for Breast Cancer Among Chinese Females. Clinical Breast Cancer 2017; 15:15.	Not mammography negative
Elizalde 2016. Additional US or DBT after digital mammography: which one is the best combination? Acta Radiologica 57(1): 13-8	Not a screening population
Giger, M. L., et al. Automated Breast Ultrasound in Breast Cancer Screening of Women With Dense Breasts: Reader Study of Mammography-Negative and Mammography-Positive Cancers. AJR. American Journal of Roentgenology 2016; 206(6): 1341-50.	Not mammography negative
Kumar, J. U., et al. Journal of Clinical and Diagnostic Research JCDR 2017; 11(8): TC29-TC32	Mixed symptomatic/ asymptomatic women
Lee 2016. Non-mass lesions on screening breast ultrasound. Medical Ultrasonography 18(4): 446-451	Not a screening population
Malaj 2016. Synergy in combining findings from mammography and ultrasonography in detecting malignancy in women with higher density breasts and lesions over 2 cm in Albania. Wspolczesna Onkologia 2016; 20(6): 475-480	Not screening population
Ohuchi, N., et al. Sensitivity and specificity of mammography and adjunctive ultrasonography to screen for breast cancer in the Japan Strategic Anti-cancer Randomized Trial (J-START): a randomised controlled trial. Lancet 2016; 387(10016): 341-8.	Women with dense breasts not shown separately
Omidiji, O. A., et al. Breast cancer screening in a resource poor country: Ultrasound versus mammography. Ghana Medical Journal 2017; 51(1): 6-12	Women with dense breasts not shown separately
Padia 2017. Detecting Breast Cancer with a Dual-Modality Device. Diagnostics 7(1): 18	Ineligible outcome
Siu 2016. Screening for Breast Cancer: U.S. Preventive Services Task Force Recommendation Statement. Annals of Internal Medicine 164(4): 279-96	Summary of USPTF review, not primary or independent study
Vourtsis, A., et al. The performance of 3D ABUS versus HHUS in the visualisation and BI-RADS characterisation of breast lesions in a large cohort of 1,886 women. European Radiology 2017; 21: 21.	Mixed screening/ diagnostic mammograms
Zhao 2015. Limitations of mammography in the diagnosis of breast diseases compared with ultrasonography: a single-center retrospective analysis of 274 cases. European Journal of Medical Research 20: 49	Not a screening population

Question 4

Study	Reason for exclusion
Abbey, C. K. 9787. A Utility/Cost Analysis of Breast Cancer Risk Prediction Algorithms	Not ultrasound
Blue Cross Blue Shield, Association 2014. Special report: screening asymptomatic women with dense breasts and normal mammograms for breast cancer. Technology Evaluation Center Assessment Program. Executive Summary 2014; 28(15): 1-2.	Not available
Bowles 2016. The Use of Ultrasound in Breast Cancer Screening of Asymptomatic Women with Dense Breast Tissue: A Narrative Review	Included film and digital studies; none of the cost studies were

	in studies using digital mammograms
Brancato, B. 2007. Negligible advantages and excess costs of routine addition of breast ultrasonography to mammography in dense breasts	Not a screening population
Corsetti 2006. Role of ultrasonography in detecting mammographically occult breast carcinoma in women with dense breasts	A subset of the women in Corsetti 2008
Corsetti 2008. Breast screening with ultrasound in women with mammography-negative dense breasts: Evidence on incremental cancer detection and false positives, and associated cost. European Journal of Cancer 2008; 44(4): 539-544	Not stated to be digital mammography (Corsetti 2011 paper states they used film 2001-2006)
De Felice, C. 2007. Diagnostic utility of combined ultrasonography and mammography in the evaluation of women with mammographically dense breasts	Film not digital mammography
Duffy 2017. Addition of ultrasound to mammography in the case of dense breast tissue: Systematic review and meta analysis	Systematic review/ meta-analysis not cost-effectiveness study
Freer 2015. Breast cancer screening in the era of density notification legislation: summary of 2014 Massachusetts experience and suggestion of an evidence-based management algorithm by multi-disciplinary expert panel	Not cost-effectiveness
Gartlehner 2013. Adjunct ultrasonography for breast cancer screening in women at average risk: A systematic review	Systematic review but the authors found no studies that met their inclusion criteria
Giuliano, V. 2013 Volumetric breast ultrasound as a screening modality in mammographically dense breasts	Duplicate
Hooley 2012. Screening US in patients with mammographically dense breasts: Initial experience with Connecticut public act 09-41. Radiology 2012; 265(1): 59-69	Mixed diagnostic/ screening population
Merry 2014. Update on Screening Breast Ultrasonography. Radiologic Clinics of North America 2014; 52(3): 527-537.	Not cost-effectiveness
Sobotka, J. 2015. Breast Density Legislation: Discussion of Patient Utilization and Subsequent Direct Financial Ramifications for Insurance Providers	Not full economic evaluation
Venturini 2013. Tailored breast cancer screening program with microdose mammography, US, and MR Imaging: short-term results of a pilot study in 40-49-year-old women. Radiology 2013; 268(2): 347-55.	Women with dense breasts not shown separately

Appendix 4 Data extraction form and tables with quality assessment

Data extraction template for questions 1, 2 and 3

Ultrasound as an add-on test after negative mammography screening in women with dense breasts

DATA EXTRACTION FORM

Review Details	
Reviewer	

Study details	
Citations for all linked publications from the same study/cohort	
First author surname (main paper for the study)	
Year of publication (main paper for the study) **(NB 2000 on for Q1/2; 2005 on for Q3/4)**	
Study/cohort name/ identifier	
Country	
Study design	
Study setting	
Number of centres	
Total study duration *(including length of follow up if applicable)*	
Funding *(government/private/ manufacturer/ other - specify)*	
Competing interests / Role of sponsor	

Aim of the study

Methods of the study	
Recruitment dates	
Inclusion criteria	
Exclusion criteria	
Recruitment method (e.g. consecutive participants)	
Statistical methods	

Baseline characteristics of women
General description of sample:

	Whole sample	Subgroup 1 (specify)	Subgroup 2 (specify)
Enrolled			
Excluded pre-baseline (plus reason)			
Sample size included at baseline (NB >100 for Q1)			
Excluded from analysis (baseline minus analysed), plus reason			
Sample size analysed			
Age (mean; SD or range)			
BMI (mean; SD or range)			
Ethnicity			
Menopausal status			
Comments on differences between study arms:			

Density measures: Q1/Q2			
	Measure 1	**Measure 2**	**Measure 3**
Density measure(s) used (name/description/version number):			
Does mammographic density measure use oblique or cranio-caudal view?			
Does the density measure use texture analysis?			
Density classifications (with description): n (%) in each group			
Comparison Q1/2: density measure 1 vs. density measure 2, or left vs. right breast etc.			
General description of raters:			
	Whole sample	**Subgroup 1**	**Subgroup 2**
Age (mean; measure of deviation)			
Profession			
Experience			
Raters blinded?			
Comments on differences between study arms:			

Interventions and comparators: Q3: mammography + ultrasound vs. mammography		
	Mammography **NB: mammography must be digital not film. State whether CR (computed radiography) or DR (digital radiography)**	**Mammography + ultrasound** **NB: describe whether ultrasound is** **A) automated:** **A i) included in the mammography plate or** **A ii) a separate machine; or** **B) handheld (must include whole breast).**

		State whether a high frequency probe was used; must be > 5MHz
Description of index test /comparator:		
1 or 2 screeners		
Experience of the operators		
Whether CAD was used or not (if automated)		
Quality of the ultrasound / mammogram		
Number receiving index test/comparator (%)		
Reference standard used		
Number receiving reference standard (%)		
Follow up (years)		

Results: Question 1: What are the test-retest and inter-rater reliability of available methods to measure mammographic breast density? What is the concordance between different methods of measuring mammographic breast density?

Inter-rater reliability

(SPECIFY MEASURE)	**Reader 2**					
Reader 1	**Category 1 (specify)**	**Category 2 (specify)**	**Category 3 (specify)**	**Category 4 (specify)**	**Total**	**Test statistic**
Category 1 (specify)						
Category 2 (specify)						
Category 3 (specify)						
Category 4 (specify)						
Total						

ADD MORE (AND ADAPT) TABLES AS REQUIRED

Test-retest reliability

(SPECIFY MEASURE)				
Time between assessments:				
Domain/category	**First assessment score**	**Second assessment score**	**Test statistic 1 (specify)**	**Test statistic 2 (specify)**
Category 1 (specify)				
Category 2 (specify)				
Category 3 (specify)				
Category 4 (specify)				

ADD MORE (AND ADAPT) TABLES AS REQUIRED

Concordance

	Measure 2 (specify)					Test statistic
Measure 1 (specify)	**Category 1 (specify)**	**Category 2 (specify)**	**Category 3 (specify)**	**Category 4 (specify)**	**Total**	
Category 1 (specify)						
Category 2 (specify)						
Category 3 (specify)						
Category 4 (specify)						
Total						

ADD MORE (AND ADAPT) TABLES AS REQUIRED

Results: Question 2: Is mammographic breast density a risk factor for cancers being missed during screening (false negatives/interval cancers)?

(specify density measure)							
Outcome (missed cancer, FN or interval)	**Density category**				**Odds ratio, risk ratio, absolute risk, mean difference (specify) (95% CI)**		**Covariates adjusted for**
	(specify)	**(specify)**	**(specify)**	**Total**	**Crude**	**Adjusted**	
Event (specify)							
Nonevent (specify)							
Total							

ADD MORE (AND ADAPT) TABLES AS REQUIRED

(specify density measure)							
Outcome (cancer)	**Density category**				**Odds ratio, risk ratio, absolute risk, mean difference (specify) (95% CI)**		**Covariates adjusted for**
	(specify)	**(specify)**	**(specify)**	**Total**	**Crude**	**Adjusted**	
Cancer							
No cancer							
Total							

ADD MORE (AND ADAPT) TABLES AS REQUIRED

Distribution of cancer type by risk group (for each test)

(specify density measure)	**Invasive**	**DCIS**	**Total**
Category 1 (specify)			
Category 2 (specify)			
Category 3 (specify)			
Category 4 (specify)			

ADD MORE (AND ADAPT) TABLES AS REQUIRED

Results: Question 3: What is the test accuracy of ultrasound following mammography in comparison to mammography to detect breast cancer in women with dense breasts?

Cancer Detection

	Disease positive	Disease negative	Total	
Mammography only				
Screening test (specify) positive				**(positive predictive value here)**
Screening test (specify) negative				**(negative predictive value here)**
Total				
	(sensitivity here)	**(specificity here)**		
Recall rate:				
Mammography plus Ultrasound				
Screening test (specify) positive				**(positive predictive value here)**
Screening test (specify) negative				**(negative predictive value here)**
Total				
	(sensitivity here)	**(specificity here)**		
Recall rate				

ADD MORE (AND ADAPT) TABLES AS REQUIRED

OR

Cancer Detection

	Mammography only		Mammography + ultrasound		Difference between mammography and mammography + ultrasound	
	N/Total	**Estimate (95% CI)**	**N/Total**	**Estimate (95% CI)**	**N/Total**	**Estimate (95% CI)**
Sensitivity						
Specificity						
PPV						
NPV						
Recall rate						

ADD MORE (AND ADAPT) TABLES AS REQUIRED

Characteristics of extra cancers detected by US only and mammography only

	Cancers detected by mammography only	Cancers detected by mammography plus ultrasound only	All screen detected cancers
Number of participants			
Number screened			
Number of cancers			
Number of invasive cancers			
Number of DCIS			
Invasive cancer grade			
High			
Intermediate			

Low			
Unknown			
DCIS grade			
High			
Intermediate			
Low			
Unknown			
Tumour size, mm (mean; SD or range)			
Stage			
No. of stage 0 cancers			
No. of stage IA or IB cancers			
No. of stage IIA or IIB cancers			
No. of stage IIIA, IIB, or IIIC cancers			
No. of stage IV cancers			
No. of unknown cancers			
ER/PR status			
ER+/PR+			
ER+/PR-			
ER-/PR-			
ER-/PR+			
Lymph node status			
Positive			
Negative			
Unknown			
HER2			
Positive			
Negative			
Unknown			
Breast density			
Category 1 (specify)			
Category 2 (specify)			
Category 3 (specify)			
Category 4 (specify)			
Immunophenotype			
Luminal A			
Luminal B			
Basal-like			
Unclassified			
Unknown			

ADD MORE (AND ADAPT) TABLES AS REQUIRED

Results: Question 4: For women attending breast screening in the UK, what are the cost-consequences of adding density measurements, and then ultrasound for those found to have high mammographic breast density?			
	Mammography	**Density measurement + ultrasound**	**p value**
Time taken for screening process (minutes)			
Cost per extra case detected **Cost per extra case detected by type (invasive/nodal involvement etc)**			

Conclusions/limitations	
Study author conclusions	
Limitations noted by the study authors	
Reviewer notes	
Abbreviations	BI-RADS: Breast Imaging-Reporting and Data System

Data extraction table for question 4

Table a. Characteristics and findings of cost-effectiveness studies investigating supplemental ultrasound in women with mammography-negative dense breasts

Author (Year)	Type of economic evaluation & model	Population studied	Comparators	Methods (perspective, time horizon and discount rate)	Methods (costs, outcomes, ICER and sensitivity analyses)

Appendix 5 Quality assessment tools

Question 1: Quality Appraisal of Diagnostic Reliability (QAREL) Checklist

Item	Yes	No	Unclear	N/A
1. Was the test evaluated in a sample of subjects who were representative of those to whom the authors intended the results to be applied?				
2. Was the test performed by raters who were representative of those to whom the authors intended the results to be applied?				
3. Were raters blinded to the findings of other raters during the study?				
4. Were raters blinded to their own prior findings of the test under evaluation?				
5. Were raters blinded to the results of the reference standard for the target disorder (or variable) being evaluated?				
6. Were raters blinded to clinical information that was not intended to be provided as part of the testing procedure or study design?				
7. Were raters blinded to additional cues that were not part of the test?				
8. Was the order of examination varied?				
9. Was the time interval between repeated measurements compatible with the stability (or theoretical stability) of the variable being measured?*				
10. Was the test applied correctly and interpreted appropriately?				
11. Were appropriate statistical measures of agreement used?**				
Total				

* <2 years

** Acceptable: Bland-Altman, ICC (for continuous data), kappa (for categorical/ordinal data – should be weighted, with an explanation of what weights were applied). Unacceptable: correlation coefficients on their own, significance testing of differences between coefficients.

Good-quality diagnostic reliability studies used a representative sample of subjects and raters, had blinded assessment of the reference standard (where applicable) and also blinded raters to non-clinical cues and to others ratings, used a varied examination order, an appropriate time interval between repeated measures, appropriate approaches to application and interpretation of the test, and used appropriate statistical measures of agreement. Diagnostic reliability studies were downgraded to fair if they were unable to meet the majority of good-quality criteria.

Question 2: QUIPS

Quality assessment - Quality in Prognostic Studies (QUIPS) tool				
Biases	**Issues to consider for judging overall rating of risk of bias**	**Study methods & comments**	**Rating of reporting**	**Rating of risk of bias**

Instructions to assess the risk of each potential bias:	***These issues will guide your thinking and judgment about the overall risk of bias within each of the 6 domains. Some 'issues' may not be relevant to the specific study or the review research question. These issues are taken together to inform the overall judgment of potential bias for each of the 6 domains.***	***Provide comments or text excerpts in the white boxes below, as necessary, to facilitate the consensus process that will follow***	***Yes, partial, no or unsure.***	***High, Moderate, or Low for 6 domains***
1. Study Participation	*Goal: To judge the risk of selection bias (likelihood that relationship between PF and outcome is different for participants and eligible non-participants).*			
Source of target population	The source population or population of interest is adequately described			
Method used to identify population	The sampling frame and recruitment are adequately described, including methods to identify the sample sufficient to limit potential bias (number and type used, e.g., referral patterns in health care)			
Recruitment period	Period of recruitment is adequately described			
Place of recruitment	Place of recruitment (setting and geographic location) are adequately described			
Inclusion and exclusion criteria	Inclusion and exclusion criteria adequately described (e.g. including explicit diagnostic criteria or zero time description)			
Adequate study participation	There is adequate participation in the study by eligible individuals			
Baseline characteristics	The baseline study sample (i.e., individuals entering the study) is adequately described			
Summary Study participation	The study sample represents the population of interest on key characteristics, sufficient to limit potential bias of the observed relationship between PF and outcome.			
2. Study Attrition	*Goal: To judge the risk of attrition bias (likelihood that relationship between PF and outcome are different for completing and non-completing participants).*			
Proportion of baseline sample available for analysis	Response rate (i.e., proportion of study sample completing the study and providing outcome data) is adequate.			
Attempts to collect information on participants who dropped out	Attempts to collect information on participants who dropped out of the study are described.			
Reasons and potential impact	Reasons for loss to follow-up are provided.			

of subjects lost to follow-up				
Outcome and prognostic factor information on those lost to follow-up	Participants lost to follow-up are adequately described There are no important differences between participants who completed the study and those who did not.			
Study Attrition Summary	Loss to follow-up (from baseline sample to study population analyzed) is not associated with key characteristics (i.e., the study data adequately represent the sample) sufficient to limit potential bias to the observed relationship between PF and outcome.			
3. Prognostic Factor Measurement	*Goal: To judge the risk of measurement bias related to how PF was measured (differential measurement of PF related to the level of outcome).*			
Definition of the PF	A clear definition or description of 'PF' is provided (e.g., including dose, level, duration of exposure, and clear specification of the method of measurement)			
Valid and Reliable Measurement of PF	Method of PF measurement is adequately valid and reliable to limit misclassification bias (e.g., may include relevant outside sources of information on measurement properties, also characteristics, such as blind measurement and limited reliance on recall). Continuous variables are reported or appropriate cut-points (i.e., not data-dependent) are used.			
Method and Setting of PF Measurement	The method and setting of measurement of PF is the same for all study participants.			
Proportion of data on PF available for analysis	Adequate proportion of the study sample has complete data for PF variable.			
Method used for missing data	Appropriate methods of imputation are used for missing 'PF' data			
PF Measurement Summary	PF is adequately measured in study participants to sufficiently limit potential bias.			
4. Outcome Measurement	*Goal: To judge the risk of bias related to the measurement of outcome (differential measurement of outcome related to the baseline level of PF).*			
Definition of the Outcome	A clear definition of outcome is provided, including duration of follow-up and level and extent of the outcome construct.			

Valid and Reliable Measurement of Outcome	The method of outcome measurement used is adequately valid and reliable to limit misclassification bias (e.g., may include relevant outside sources of information on measurement properties, also characteristics, such as blind measurement and confirmation of outcome with valid and reliable test).			
Method and Setting of Outcome Measurement	The method and setting of outcome measurement is the same for all study participants.			
Outcome Measurement Summary	Outcome of interest is adequately measured in study participants to sufficiently limit potential bias			
5. Study Confounding	*Goal: To judge the risk of bias due to confounding (i.e. the effect of PF is distorted by another factor that is related to PF and outcome).*			
Important Confounders Measured	All important confounders, including treatments are measured.			
Definition of the confounding factor	Clear definitions of the important confounders measured are provided (e.g., including dose, level, and duration of exposures).			
Valid and Reliable Measurement of Confounders	Measurement of all important confounders is adequately valid and reliable (e.g., may include relevant outside sources of information on measurement properties, also characteristics, such as blind measurement and limited reliance on recall)			
Method and Setting of Confounding Measurement	The method and setting of confounding measurement are the same for all study participants			
Method used for missing data	Appropriate methods are used if imputation is used for missing confounder data			
Appropriate Accounting for Confounding	Important potential confounders are accounted for in the study design (e.g., matching for key variables, stratification, or initial assembly of comparable groups) Important potential confounders are accounted for in the analysis (i.e., appropriate adjustment)			
Study Confounding Summary	Important potential confounders are appropriately accounted for, limiting potential bias with respect to the relationship between PF and outcome.			

6. Statistical Analysis and Reporting	*Goal: To judge the risk of bias related to the statistical analysis and presentation of results*			
Presentation of analytical strategy	There is sufficient presentation of data to assess the adequacy of the analysis			
Model development strategy	The strategy for model building (i.e., inclusion of variables in the statistical model) is appropriate and is based on a conceptual framework or model. The selected statistical model is adequate for the design of the study			
Reporting of results	There is no selective reporting of results.			
Statistical Analysis and Presentation Summary	The statistical analysis is appropriate for the design of the study, limiting potential for presentation of invalid or spurious results			

Question 3:

USPTF criteria for assessing internal validity of individual diagnostic accuracy studies

Criteria:	**Notes for completion of assessment**	**Adequatein this study? Yes/No/Unsure/N/A (Yes = a good quality outcome)**
Screening test relevant, available for primary care, and adequately described	Screening test = Digital mammography; HHUS or ABUS (whole breast)	
Credible reference standard, performed regardless of test results	Reference standard = Biopsy/histology result for breast cancer; follow up for at least 1 year for interval cancers/true negatives	
Reference standard interpreted independently of screening test	Requires follow up for interval cancers, not just histology/biopsy	
Indeterminate results handled in a reasonable manner	Short term repeat exams are OK	
Spectrum of patients included in study	Must be a screening population; OK to include or exclude prior breast cancer, high risk women, prior breast surgery as part of the population (but population must not be exclusively high risk, symptomatic, or diagnostic)	
Sample size	No minimum sample size but quality downgraded if <100 people	

Reliable screening test	Mammography and ultrasound can be assumed reliable in this context; excludes untested experimental methods	
Global rating of internal validity		

Definition of ratings based on above criteria:
Good: Evaluates relevant available screening test; uses a credible reference standard; interprets reference standard independently of screening test; assesses reliability of test; has few or handles indeterminate results in a reasonable manner; includes large number (>100) of broad-spectrum patients with and without disease
Fair: Evaluates relevant available screening test; uses reasonable although not best standard; interprets reference standard independent of screening test; has moderate sample size (50 to 100 subjects) and a "medium" spectrum of patients
Poor: Has a fatal flaw, such as: Uses inappropriate reference standard; improperly administers screening test; biased ascertainment of reference standard; has very small sample size or very narrow selected spectrum of patients

USPTF criteria for assessing external validity (generalizability) of individual studies

Each study that is identified as providing evidence to answer a key question is assessed according to its external validity (generalizability), using the following criteria.

Criteria:	**Notes for completion of assessment**	**Adequate in this study? Yes/No/Unsure/N/A (Yes = a good quality outcome so all items are scored in the same direction)**
Study population: The degree to which a study's subjects constitute a special population—either because they were selected from a larger eligible population or because they do not represent persons who are likely to seek or be candidates for the preventive service.		
Demographic characteristics (i.e., age, sex, ethnicity, education, income): The criteria for inclusion/exclusion or nonparticipation do not encompass the range of persons who are likely to be candidates for the preventive service in the U.S. primary care population.	Must include majority of women in age range 50-70; downgrade if >50% outside this age range	
Comorbid conditions: The frequency of comorbid conditions in the study population does not represent the frequency likely to be encountered in persons who seek the preventive service in the U.S. primary care population.	Downgrade if majority high risk women	
Special inclusion/exclusion criteria: There are other special inclusion/exclusion criteria that make the study population not representative of the U.S. primary care population.	Flag up ethnicity	
Refusal rate (i.e., ratio of included to not included but eligible participants): The refusal rate among eligible study subjects is high, making the study population not representative of the	Downgrade if refusal rate >10%	

U.S. primary care population, even among eligible enrollees.		
Adherence (i.e., run-in phase, frequent contact to monitor adherence): The study design has features that may increase the effect of the intervention in the study more than would be expected in a clinically observed population.	Flag up screening interval (UK = 3 years)	
Stage or severity of disease: The selection of subjects for the study includes persons at a disease stage that is earlier or later than would be found in persons who are candidates for the preventive service.	Should be a general screening sample: OK to include or exclude prior breast cancer, high risk women, prior breast surgery as part of the population (but population must not be exclusively high risk, symptomatic, or diagnostic)	
Recruitment: The sources for recruiting subjects for the study and/or the effort and intensity of recruitment may distort the characteristics of the study subjects in ways that could increase the effect of the intervention as it is observed in the study.	Should be general screening population	
Study setting: The degree to which the clinical experience in the setting in which the study was conducted is likely to be reproduced in other settings:		
Health care system: The clinical experience in the system in which the study was conducted is not likely to be the same as that experienced in other systems (e.g., the system provides essential services for free when these services are only available at a high cost in other systems).	Universal screening programme or selected	
Country: The clinical experience in the country in which the study was conducted is not likely to be the same as that in the United States (e.g., services available in the United States are not widely available in the other country or vice versa).	Flag up country	
Selection of participating centers: The clinical experience in which the study was conducted is not likely to be the same as in offices/hospitals/settings where the service is delivered to the U.S. primary care population (e.g., the center provides ancillary services that are not generally available).	General screening programme or tertiary centre where problematic cases referred in	
Time, effort, and system cost for the intervention: The time, effort, and cost to develop the service in the study is more than would be available outside the study setting.	Should be a routine screening service	
Study providers: The degree to which the providers in the study have the skills and expertise likely to be available in general settings:		

Training to implement the intervention: Providers in the study are given special training not likely to be available or required in U.S. primary care settings.	Should be general screening service not unusually highly trained operators	
Expertise or skill to implement the intervention: Providers in the study have expertise and/or skills at a higher level than would likely be encountered in typical settings.	Should be general screening service not unusually highly skilled operators	
Ancillary providers: The study intervention relies on ancillary providers who are not likely to be available in typical settings.	Should be radiologists/radiographers	
Global rating of external validity		

USPTF Global rating of external validity (generalisability)

External validity is rated "good" if:
• The study differs minimally from the U.S. primary care population/setting/providers and only in ways that are unlikely to affect the outcome; it is highly probable (>90%) that the clinical experience with the intervention observed in the study will be attained in the U.S. primary care setting.

External validity is rated "fair" if:
• The study differs from the U.S. primary care population/setting/providers in a few ways that have the potential to affect the outcome in a clinically important way; it is moderately probable (50% to 89%) that the clinical experience with the intervention observed in the study will be attained in the U.S. primary care setting.

External validity is rated "poor" if:
• The study differs from the U.S. primary care population/setting/providers in many ways that have a high likelihood of affecting the clinical outcome; probability is low (<50%) that the clinical experience with the intervention observed in the study will be attained in the U.S. primary care setting.

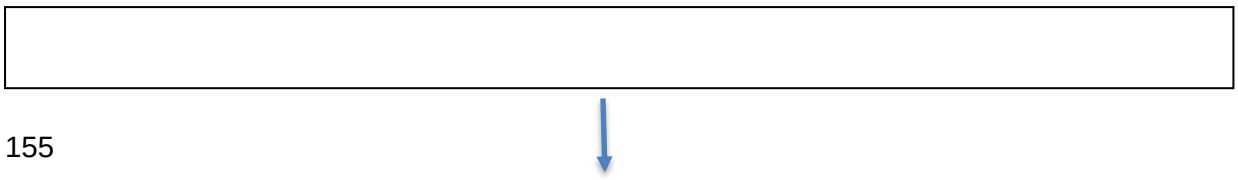
QUADAS-2 (adjusted)

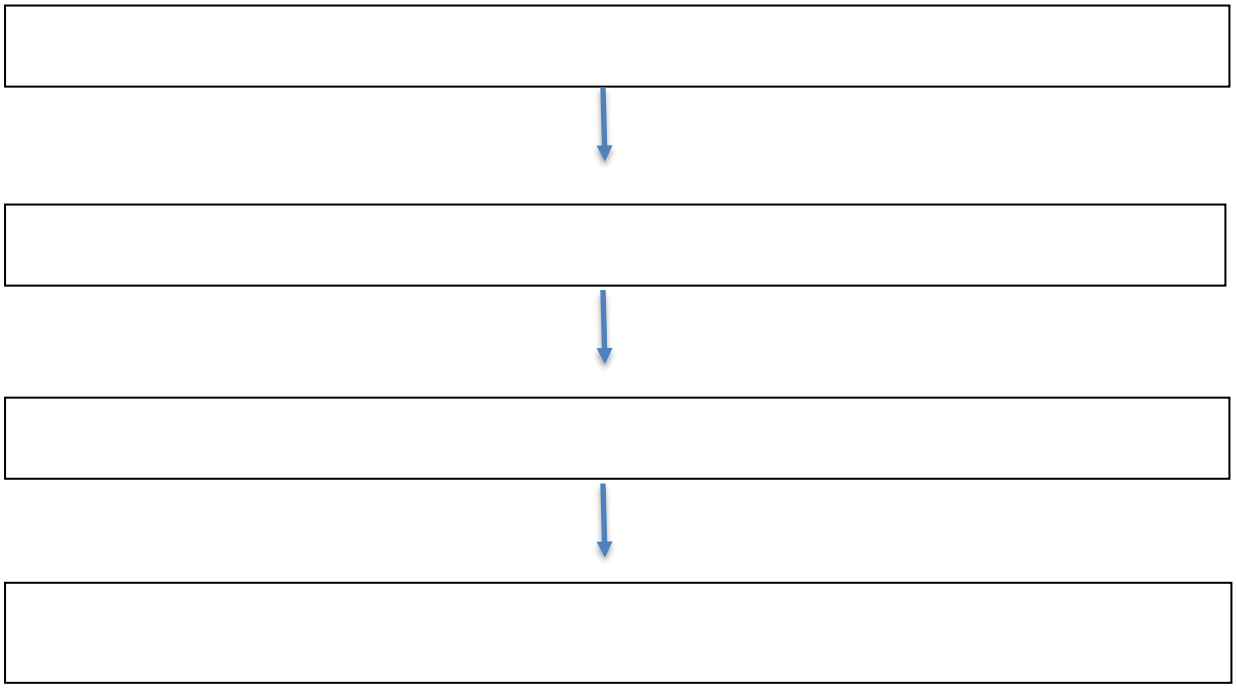
First author surname and year of publication:

Name of first reviewer: ***Name of second reviewer:***
Phase 1: State the review question:
What is the test accuracy of ultrasound following mammography in comparison to mammography to detect cancer in women with dense breasts?

Patients (setting, intended use of index test, presentation, prior testing): women with mammographically normal, but dense breasts
Index test(s): Ultrasound
Reference standard and target condition: Biopsy/histology for cancer; follow up for at least 1 year for negative screen

Phase 2: Draw a flow diagram for the primary study





Phase 3: Risk of bias and applicability judgments

QUADAS-2 is structured so that 4 key domains are each rated in terms of the risk of bias and the concern regarding applicability to the research question (as defined above). Each key domain has a set of signalling questions to help reach the judgments regarding bias and applicability.

DOMAIN 1: PATIENT SELECTION **A. Risk of Bias**
Describe methods of patient selection:
+ Was a consecutive or random sample of patients enrolled? Yes/No/Unclear **+ Was a consecutive or random sample of women who screened negative AND had dense breasts followed up with ultrasound? Yes/No/Unclear** **+ Was a case-control design avoided? Yes/No/Unclear** **+ Did the study avoid inappropriate exclusions? Yes/No/Unclear** **Could the selection of patients have introduced bias? RISK: LOW/HIGH/UNCLEAR**
B. Concerns regarding applicability
Describe included patients (prior testing, presentation, intended use of index test and setting):
Is there concern that the included patients do not match the review question? CONCERN: LOW/HIGH/UNCLEAR

DOMAIN 2: INDEX TEST (mammography)

If more than one index test was used, please complete for each test. **A. Risk of Bias**	
Describe the index test and how it was conducted and interpreted:	
+ Were the index test results interpreted without knowledge of the results of the reference standard? **Could the conduct of the index test have introduced bias?**	**Yes/No/Unclear** **RISK: LOW/HIGH/UNCLEAR**
B. Concerns regarding applicability **Is there concern that the index test, its conduct, or interpretation differ from the review question?**	**CONCERN: LOW/HIGH/UNCLEAR**

DOMAIN 2: INDEX TEST (ultrasound) **If more than one index test was used, please complete for each test.** **A. Risk of Bias**	
Describe the index test and how it was conducted and interpreted:	
+ Were the index test results interpreted without knowledge of the results of the reference standard? **Could the conduct of the index test have introduced bias?**	**Yes/No/Unclear** **RISK: LOW/HIGH/UNCLEAR**
B. Concerns regarding applicability **Is there concern that the index test, its conduct, or interpretation differ from the review question?**	**CONCERN: LOW/HIGH/UNCLEAR**

DOMAIN 3: REFERENCE STANDARD **A. Risk of Bias**	
Describe the reference standard and how it was conducted and interpreted:	
+ Is the reference standard likely to correctly classify the target condition? **+ Were the reference standard results interpreted without knowledge of the results of the index test?** **Could the reference standard, its conduct, or its interpretation have introduced bias?**	**Yes/No/Unclear** **Yes/No/Unclear** **RISK: LOW/HIGH/UNCLEAR**
B. Concerns regarding applicability **Is there concern that the target condition as defined by the reference standard does not match the review question?**	**CONCERN: LOW/HIGH/UNCLEAR**

DOMAIN 4: FLOW AND TIMING **A. Risk of Bias**

Describe any patients who did not receive the index test(s) and/or reference standard or who were excluded from the 2x2 table (refer to flow diagram):	
Describe the time interval and any intervention between (1) the two index tests (mammography versus ultrasound) and (2) the index tests(s) and reference standard:	
+ Was there an appropriate interval between the two index tests?	**Yes/No/Unclear**
+ Was there an appropriate interval between index test(s) and reference standard?	**Yes/No/Unclear**
+ Did all patients receive a reference standard?	**Yes/No/Unclear**
+ Did all patients receive the same reference standard?	**Yes/No/Unclear**
+ Were all patients included in the analysis?	**Yes/No/Unclear**
Could the patient flow have introduced bias?	**RISK: LOW/HIGH/UNCLEAR**

Question 4: CHEERS

Critical appraisal of the economic evaluation studies using the CHEERS checklist (each column = 1 study)

CHEERS checklist[33]								
Title and abstract								
1 Title: Identify the study as an economic evaluation, or use more specific terms such as ``cost-effectiveness analysis``, and describe the interventions compared.								
2 Abstract: Provide a structured summary of objectives, methods including study design and inputs, results including base case and uncertainty analyses, and conclusions.								
Introduction								
3 Background & objectives: Provide an explicit statement of the broader context for the study. Present the study question and its relevance for health policy or practice decisions.								
Methods								
4 Target Population and Subgroups: Describe characteristics of the base case population and subgroups analysed including why they were chosen.								
5 Setting and Location: State relevant aspects of the system(s) in which the decision(s) need(s) to be made.								

6 Study perspective: Describe the perspective of the study and relate this to the costs being evaluated.								
7 Comparators: Describe the interventions or strategies being compared and state why they were chosen.								
8 Time Horizon: State the time horizon(s) over which costs and consequences are being evaluated and say why appropriate.								
9 Discount Rate: Report the choice of discount rate(s) used for costs and outcomes and say why appropriate.								
10 Choice of Health Outcomes: Describe what outcomes were used as the measure(s) of benefit in the evaluation and their relevance for the type of analysis performed.								
11a Measurement of Effectiveness - Single Study-Based Estimates: Describe fully the design features of the single effectiveness study and why the single study was a sufficient source of clinical effectiveness data.								
11b Measurement of Effectiveness - Synthesis-based Estimates: Describe fully the methods used for identification of included studies and clinical effectiveness data synthesis of clinical effectiveness data.								
12 Measurement and Valuation of Preference-based Outcomes: If applicable, describe the population and methods used to elicit preferences for health outcomes.								
13a Estimating Resources and Costs - Single Study-based Economic evaluation: Describe approaches used to estimate resource use associated with the alternative interventions. Describe primary or secondary research methods for valuing each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs.								
13b Estimating Resources and Costs - Model-based Economic Evaluation: Describe approaches and data sources used to estimate resource use associated with model health states. Describe primary or secondary research methods for valuing each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs.								
14 Currency, Price Date and Conversion: Report the dates of the estimated resource quantities and unit costs. Describe methods for								

adjusting estimated unit costs to the year of reported costs if necessary. Describe methods for converting costs into a common currency base and the exchange rate.								
15 Choice of Model: Describe and give reasons for the specific type of decision-analytic model used. Providing a figure to show model structure is strongly recommended.								
16 Assumptions: Describe all structural or other assumptions underpinning the decision-analytic model.								
17 Analytic Methods: Describe all analytic methods supporting the evaluation. This could include methods for dealing with skewed, missing or censored data, extrapolation methods, methods for pooling data, approaches to validate a model, & methods for handling population heterogeneity and uncertainty.								
Results								
18 Study parameters: Report the values, ranges, references, and if used, probability distributions for all parameters. Report reasons or sources for distributions used to represent uncertainty where appropriate. We strongly recommend the use of a table to show the input values.								
19. Incremental costs and outcomes: For each intervention, report mean values for the main categories of estimated costs and outcomes of interest, as well as mean differences between the comparator groups. If applicable, report incremental cost-effectiveness ratios.								
20a Characterizing Uncertainty - Single study-based economic evaluation: Describe the effects of sampling uncertainty for the estimated incremental cost and incremental effectiveness, parameters together with the impact of methodological assumptions.								
20b Characterizing Uncertainty - Model-based economic evaluation: Describe the effects on the results of uncertainty for all input parameters, and uncertainty related to the structure of the model and assumptions.								
21 Characterizing Heterogeneity: If applicable, report differences in costs, outcomes or in cost-effectiveness that can be explained by variations between subgroups of patients with different baseline characteristics or other observed variability in effects that are not reducible by more information.								

Discussion								
22 Study Findings, Limitations, Generalizability, and Current Knowledge: Summarize key study findings and describe how they support the conclusions reached. Discuss limitations and the generalizability of the findings and how the findings fit with current knowledge.								
Other								
23 Source of Funding: Describe how the study was funded and the role of the funder in the identification, design, conduct and reporting of the analysis. Describe other non-monetary sources of support.								
24 Conflicts of Interest: Describe any potential for conflict of interest among study contributors in accordance with journal policy. In the absence of a journal policy, we recommend authors comply with International Committee of Medical Journal Editors' recommendations								

Key: y = yes, n = no, N/A = not applicable and * = partially completed

Appendix 6 Included studies

Question 1

Table a: Design and quality issues

	Study	Population (n)	Interventions/ Comparator	Outcome	No. centres; country	Quality summary	Sample rep.?	Readers rep.?	Time <2 years?	Limitations
1.	Abdolell 2013[37]	Digital mammograms – no further information (n=138)	Densitas and visual percent density assessment	Inter-rater reliability; concordance between Densitas and visual assessment	1; Canada	Fair	Unclear	Yes	Unclear	The Pearson correlation coefficient (ρ) provides an inadequate, inflated, and overoptimistic measure of the level of agreement. This measure is not eligible for our review.
2.	Alshafeiy 2017[48]	Consecutive women undergoing screening with digital 2D mammography and tomosynthesis with a negative or benign (category 1 and 2) outcome (n=309); mean (SD) age 65.7 ± 11.4 years (range, 35–93 years).	BI-RADS 5th edition from digital 2D images	Interreader agreement	1; USA	Fair	No	Yes	Yes	Relatively small number of readers from a single institution; results may differ in a larger study with more readers. No reference standard for breast density

3.	Conant 2017[17]	Women with 2D bilateral MLO view sDM and standard dose "For presentation" DM images available (3668 women with 7336 MLO images)	BIRADS 5th edition; LIBRA algorithm in DM	Analysis of variance to determine whether the automated percent density estimates for DM varied significantly according to the corresponding BIRADS breast density categories	1; USA	Fair	No	No	N/A	A single area-based density estimation method using data from a single institution
4.	Destounis 2017[18]	Women diagnosed with cancer within the screening programme; mean (SD) age 62.1 (11) (n=595)	BIRADS 4th edition, from previous normal mammogram vs. Volpara v1.4.2 from previous normal mammogram if raw images available or contralateral breast if raw images not available	Agreement between visual BIRADS and automated density grade	1; USA	Fair	No	Unclear	Yes	Interval cancers not differentiated between true interval, missed or mammographically occult (i.e. masked by dense tissue).
5.	Ekpo 2016.[36]	Women who underwent DBT investigation in 2015 and had a prior DM obtained in 2014 (n=234)	BI-RADS 5th edition	BI-RADS 5th edition inter-reader reproducibility	1; Australia	Fair	No	Yes	Yes	The proportion of BIRADS D density category in the dataset is higher than that of a typical population distribution, as women that have DBT subsequent to

										DM are more likely to have dense breast than fatty breasts. No agreed standard for BD assessment.
6.	Ekpo 2016.[19]	Females who underwent screening mammography between March and July 2014 (n=292)	Quantra 2.0 vs. BIRADS 4th edition	Agreement between each radiologist and the majority report. Inter-reader agreement was assessed by comparing the first assessment of the radiologists in pairs. Intra-reader agreement was assessed by comparing the first and second readings of each radiologist.	1; Australia	Good	Unclear	Yes	Yes	The high level of agreement between the 6 radiologists may be due to the readers all working in the same practice; it is possible they would demonstrate considerable inter-reader variability with readers from different practice, limiting generalizability. Using the majority report in Phase 1 might have been a better reference standard. It is possible that the increased sensitivity of Quantra for BIRADS 1 and 2 in Phase 2 may be due to the small sample size compared with Phase 1 and the laboratory effect.
7.	Eng 2014[9] and Busana 2016[53]*	Cases: women with newly diagnosed	BI-RADS 4th edition; Cumulus	Inter- and intra-method and left-	2; UK	Good	No	Yes	Yes	The study population was

		breast cancer (mean (SD) age: 67.5 (12.7) years; not eligible as diagnostic population); controls: women who attended routine screening and were found to be breast cancer free (mean (SD) age: 59.5 (6.6) years) (n=1969)	v3; ImageJ-based method; Volpara v1.0; Quantra v1.3; single energy x-ray absorptiometry (SXA) method, v6.5	right comparisons among controls. Within-observer reliability of Cumulus. Between-observer reliability of Cumulus. LIBRA						predominantly postmenopausal, thus, limiting the generalizability of the findings to premenopausal women. Response rates were low for healthy controls (51%). Processed images were missing for 15 % of the control participants due to a logistical error.
8.	Eom 2017[45]	Healthy women (n=1000)	BIRADS 5th edition, Volpara version 1.5.12	Intra- and inter-reader agreement for BIRADS; concordance between Volpara and BIRADS	1; Republic of Korea	Good	100% Asian	Unclear	Yes	First, all the mammographic examinations were performed in a single mammographic unit, with only one specific kind of automated quantitative measurement to be used for comparisons. However, employing a unified equipment and software might have increased the data reliability. Second, the number of the readers was

										small and they were all trained at the same institution. However, we tried to assess the differences between the readers with different experience levels, which would reflect the situation often found in clinical practice. Finally, the automated volumetric measurement was used as a reference standard. The revised fifth edition of BI-RADS no longer indicates the ranges of the percentage of dense tissue and emphasizes the changes in mammography sensitivity. There is no other standard reference for mammographic density assignment in clinical practice.
9.	Garrido-Estepa 2010[46]	Women aged ≥4 years who attended screening	BI-RADS 4th edition	Intra-observer reliability	3; Spain	Fair	Unclear	No	Yes	1 reader only.

		in Barcelona, Burgos, Corunna (Coruña), Palma de Mallorca, Pamplona, Valencia and Zaragoza (n=1532)								
10.	Gweon 2013[42]	Full-field digital mammography (FFDM) examinations (n= 778)	BIRADS 4th edition; Volpara version 1.5.1	Inter-rater reliability for BIRADS. Concordance between BIRADS and Volpara	1; South Korea	Fair	Unclear	Yes	Yes	A reference standard to evaluate breast density does not exist. Three radiologists in a single institution assigned BI-RADS density categories. It would be best to perform a larger study with more patients and radiologists from a variety of practice settings to validate the findings.
11.	Harvey 2013[49]	Women aged ≥ 40 years who underwent ≥2 digital screening mammography examinations <36 months apart; mean (SD) age 57.7 +/- 11.4 (range 40-89 or older) years (n=87066)	BIRADS 3rd edition (prior to 2003) or 4th edition (released in 2003)	BIRADS test-retest agreement	5; USA	Fair	Yes	Yes	Yes	Included density interpretations determined on both 3rd and 4th editions of BIRADS lexicon

12.	Holland 2016[40]	Women aged 50-75 with consecutive exam pairs; mean (SD) age 58.8 ± 6.7 years (n=500)	Volpara v 1.5.0 and BIRADS 4th edition	Inter-exam agreement was calculated with Cohen's weighted kappa. Intraclass correlation coefficients (ICCs) were calculated to examine the interexam agreement of the four classes categorisation.	Not stated but multiple; The Netherlands	Good	Yes	Yes	No	The readers had a minimum interval of only one week between readings (although 30 months between prior and current mammograms). It may well be that variability of their criteria for the categorisation increases with the interval length, which would cause a decrease of agreement over time. In that regard, in screening practice the reader agreement might be lower than the authors found, because the screening interval is in reality much longer than the interval in this experiment.
13.	Irshad 2016[12]	Consecutive women with digital mammograms from screening mammography database; mean	BIRADS 4th edition and BIRADS 5th edition	Each radiologist evaluated the breast density of 104 mammographic examinations four	1; USA	Good	Unclear	Yes	Yes	One limitation of the study was its design for readers to focus all their attention on breast density, making density the

		age 47 (range 36-82) years (n=104)		times: twice using the 4th-edition BI-RADS criteria and twice using the 5th-edition. Intra-reader and interreader agreements for 4th-edition and 5th-edition criteria.						most important finding on the mammograms, which is not the case in real practice in which density is usually a secondary focus of attention.
14.	Irshad 2017[51]	Digital screening mammograms read by the 5 readers at the authors' institution who had read mammograms under 4th (n= 19066) or 5th (n= 16907) edition BIRADS guidelines	BIRADS 4th edition and BIRADS 5th edition	Intraclass correlation coefficient (ICC) within each dataset.	1; USA	Fair	Yes	Yes	Yes	Single institution; practice patterns of the readers might have been more similar to one another than those seen across various institutions and practices
15.	Jeffers 2017[14]	Cases: women who underwent screening mammography and subsequently received a diagnosis of breast cancer; pre-diagnostic mammographic examination at least 1 year before the date of	Cumulus 6 (version 4.0); Volpara (version not stated) and BI-RADS (version not stated)	Correlation between methods	1; USA	Fair	Unclear	Yes	Unclear	The available sample size limited the ability to detect subtle differences in discrimination among the density assessment methods. Second, clinical BI-RADS density assessment was made by a single reader. The Cumulus

		diagnosis; image of the noncancerous breast contralateral to the affected breast (n=125; 58.4% >50 years). Controls: women without a history of breast cancer who underwent screening mammography; breast cancer–free status confirmed with at least 10 years of follow-up for women aged ≥50 years or ≥3 screening mammograms negative for cancer (BI-RADS assessment category 1 or 2) for women < 50 years (n=274; 58.8% >50 years).								assessments were performed by a single reader. The standard of practice for using Cumulus software is to require the reader to undergo specialised training and attain high levels of intrareader reproducibility with test images before reading the study images. The extensive training and time required to perform Cumulus measurements made it impractical to have more than one Cumulus reader for this study, although we acknowledge that having multiple readers could have strengthened the results.
16.	Kang 2016[43]	Craniocaudal mammograms of subjects who were involved in a breast cancer screening program	Cumulus (version 4.0)	Intra- and inter-reader reliability with Cumulus	1; South Korea	Fair	No	Yes	Yes	The authors chose readers who had sufficient experience in mammographic reading and breast density estimation,

		and found to have normal breasts; mean 50.2 years; range, 28–79 years (n=100)								the small number of readers limits the generalizability of the study findings. They used only craniocaudal mammograms. Studies have shown better associations between percentage density and breast cancer on craniocaudal images than on mediolateral oblique images. Density estimates were made on images acquired from a single model of equipment. Because each type of mammographic system has different imaging characteristics and post-processing options, our study results cannot be directly applied to mammograms obtained with other types of equipment.
17.	Kerlikowske 2017[52]	Digital screening examinations of women with	BIRADS 5th edition, Volpara version 1.5.0	Correlation between BIRADS categories and	Not stated; USA	Fair	Yes	Yes	Yes	In studies for interrater and intrarater reliability

		incident invasive breast cancers and matched control subjects without prior breast cancer. (n=5406)		Volpara continuous dense breast volume, divided into quartiles						of the BI-RADS categories, investigators have reported moderate to substantial agreement. Thus, misclassification of BI-RADS categories may have influenced our results, such that some of the differences we observed could result in an under- or overestimation of associations. Our population was predominantly white and Asian; studies should be repeated with black and Hispanic women to ensure generalizability of results across all racial/ethnic groups.
18.	Llobet 2014,[15] Martinez Gomez 2014[54] and Pollan 2013[55]	Mammograms from women participants at two screening centers equipped with full-field digital mammography machines; range	BIRADS 3rd edition, DM-Scan, Cumulus	Inter- and intra-rater concordance with DM-Scan and BIRADS. Agreement between visual scale and Cumulus versus DM-Scan, with Cumulus/DM-	2; Spain	Fair	Yes	Yes	Yes	Brightness correction could introduce a significant error in MD measurement. A hard classification scheme was used, assuming that each pixel can only belong

		45-69 years (n=655)		Scan having CCC and Bland-Altman plots.						to one of the two possible classes. The relation between MD and breast cancer risk was not tested with a soft or probabilistic classification scheme, in which each pixel has an associated probability of belonging to each class. The authors did not estimate the extra time necessary to add the estimation of breast density to daily routine. DM-Scan and Cumulus were used on processed mammograms that depend on the manufacturers; the authors did not have access to raw (unprocessed) images because Spanish screening centres discard them due to storage constraints. Reliability of DM-Scan and Cumulus

										not compared in this study.
19.	Lobbes 2012[16]	Women with digital mammograms; mean 51.6 (range 23.9-91.2) years (n=200)	BIRADS 4th edition, QWIN semi-automated thresholding	Inter-reader reliability of BIRADS 4th edition; QWIN ICC left versus right breast	1; The Netherlands	Fair	Unclear	Unclear	Yes	The study included relatively small numbers of dense breasts (BIRADS 3 or 4). A true gold standard for the assessment of breast density is lacking.
20.	Mazor 2016[39]	Patients who had undergone consecutive mammography between January and March 2014 were randomly chosen; age not stated (n=503)	BIRADS 5th edition	Inter-observer agreement between technologists and radiologists. Intra- and inter-observer agreements within the group of radiologists and the inter-observer agreement within the group of technologists.	1; Israel	Good	Unclear	Yes	Yes	The reference range for breast density used in this study stemmed from the subjective measurements performed by the radiologists, as methods of objective breast density measurement such as automated breast density measuring algorithms are unavailable in the authors' institution.
21.	Osteras 2016[41] and Osteras 2016[56]	Women with digital mammograms; mean (SD) age 59.3 (5.6) years; range 50-70 years (n=537)	BIRADS 4th edition, Quantra version 2.0 (areometric density, volumetric density, BIRADS-like categories)	Inter-observer variability for each radiologist versus the median BIRADS score (unweighted	1; Norway	Fair	Unclear	Yes	Yes	The radiologists had a range of experience from 1-34 years, but more- and less-experienced readers equally influence the

				kappa and with quadratic weights)						median score. The radiologists did not use the BIRADS density scale in their daily practice but the three categories used in the Norwegian breast cancer screening program. They trained in the use of BIRADS before the study began; the training could reduce the variation in their assessments. This is a single-centre study, using the BIRADS 4th edition, but in the future the 5th edition will be used.
22.	Raza 2016[50]	Digital bilateral screening mammograms; age not stated (n=200)	BIRADS 4th edition; Volpara version not stated	Inter-rater reliability of radiologists using BIRADS before and after training, compared with a) senior breast imagers (leads truth [LT]) and b) Volpara (quantitative truth [QT]).	1; USA	Fair	No	Yes	Unclear	There is no gold standard for breast density assessment at this time. Today's software is not yet able to account for the complexity of breast tissue, as a trained radiologist can.

23.	Sartor 2016[47]	Digital mammograms with available raw data from the Malmo Breast Tomosynthesis Screening Trial (MBTST), a prospective study comparing MLO DBT alone vs. CC and MLO DM; mean age 58 (range 40-76) years (n=8426).	BIRADS 4th edition and Volpara (version 1.5.11)	Inter-observer variability for examinations with two BIRADS scores. Kappa values for comparison between Volpara density grades (VDG; categorical variable with four groups) and BIRADS scores calculated using separate kappa coefficients for each reader vs. Volpara, then results combined in a meta-analysis, weighting them using the standard error for each kappa, rendering a pooled kappa.	1; Sweden	Fair	Unclear	Yes	Unclear	Initial trial participation rate was 71.1%; further women did not have both BIRADS and Volpara readings, so overall around 67% participation.
24.	Seo 2013[44]	Healthy women received four-view screening mammograms whose mammograms were considered to be negative (BI-RADS category 1); mean 49.1 (range	BIRADS 4th edition and Volpara (version 1.4)	Intra- and inter-observer agreement for the BI-RADS density category; concordance	1; Republic of Korea	Fair	No	Yes	Yes	There is a lack of reference-standard regarding breast density. Only a small number of radiologists read the BI-RADS breast categories. <30% of eligible women consented.

		35–72) years (n=193)								
25.	Singh 2016[38]	Asymptomatic females >35 years of age; mean (SD) 48.8 (7.07), range 36-76 years (n= 476)	BIRADS 5th edition and Volpara (version 1.4.5)	Interobserver agreement using BIRADS; correlation between BIRADS and volumetric breast density	1; India	Fair	Yes	Yes	Yes	This was a small study in a single institution and examinations were interpreted by only 2 radiologists. There is no reference standard for breast density. Factors such as BMI were not investigated. Only one mammography machine was used so results cannot be generalised to all types of machines.
26.	Sprague 2016[22]	Screening mammography; mean (SD) 57.9 (10.8), range 40 to 89 years (n= 145,123)	BI-RADS 4th edition	Inter-rater variation between radiologists; test-retest reliability when interpreted by the same or a different radiologist	30; USA	Fair	Yes	Yes	Yes	The study was limited to assessments by radiologists practicing in the clinical networks of the 3 PROSPR breast cancer screening research centers. Although these included a large number of academic and community practice breast imaging facilities in 4 states, the degree of variation in breast

										density assessment may differ in other clinical settings around the country. Variation in density assessment may differ at radiology practices serving a different demographic mix of patients. Quantitative density measures were not available for comparison with the radiologist's subjective assessment. Results likely reflect not only variation in radiologist interpretation of images but also the variation in the mammography machines and software used to produce digital mammographic images that is routinely present across and within facilities over time in clinical practice.

										Over 15% of women were excluded.
27.	van der Waal 2015[13]	Screening mammograms; median age 59 (IQR: 54–64) years (n=992)	BI-RADS 5th edition; Quantra (version 1.3); Volpara (version 1.5.11)	Intra- and inter-rater reliability of the BI-RADS density scores; overall proportions of agreement (absolute agreement); intraclass correlation coefficients (ICC) between volumetric breast density estimates and BI-RADS classification	1; The Netherlands	Good	Yes	Yes	Unclear	The authors did not have any information on breast cancer risk, which would ultimately be needed to validate both breast density measures and potentially implement them in a breast cancer screening setting if they are to be used for risk stratification. More research is needed as well on the association between volumetric density and sensitivity of digital mammography. This information is required to identify a clinically relevant breast density cut-off value above which additional screening (e.g., with MRI or ultrasound) may be cost effective. Studies are also needed on the

										potential inclusion of volumetric density in risk models.

Table b: Results: Test-retest reliability

Study	Intervention	Readers	Time between assessments	Outcome reported
Ekpo 2016.[19]	BI-RADS 4th edition	All Royal Australian and New Zealand College of Radiology-certified breast radiologists. Number of years certified: R1: 13; R2: 20; R3: 3; R4: 20; R5: 19; R6: 35 (mean 18.3). Number of years reading screening mammograms: 13; 20; 3; 20; 19; 25, respectively (mean 16.7).	5 months	Weighted kappa (weighting not stated): four-category scale: Reader 1: 0.87 (0.83–0.92) Reader 2: 0.86 (0.83–0.91) Reader 3: 0.88 (0.85–0.93) Agreement between the BIRADS assessment in Phase 1 and the majority report in Phase 2 was 0.78 (0.73 to 0.85). Weighted kappa: two-category scale: Reader 1: 0.91 (0.88–0.95) Reader 2: 0.88 (0.83–0.92) Reader 3: 0.90 (0.87–0.94)
Eom 2017[45]	BIRADS 5th edition	Two were breast-imaging experts with more than five years of experience in reading mammograms, two were general radiologists with fewer years of experience in reading mammograms, and two were medical students without clinical experience in breast imaging. Two medical students were trained to read total of 80 mammogram set comprised of 20 mammograms per each Volpara density categories.	2 months	Weighted kappa (weighting not stated): Intra-reader agreement for the BI-RADS density categories a, b, c, and d was almost perfect or substantial (k=0.74–0.95) for breast-imaging experts (0.84, 0.87), general radiologists (0.86, 0.95), and students (0.74, 0.86). Intra-reader agreements on the non-dense and dense group classification were almost perfect or substantial (k=0.76–0.95) among the breast-imaging experts (0.85, 0.88), general radiologists (0.88, 0.95), and students (0.76, 0.90).

Garrido-Estepa 2010[46]	BI-RADS 4th edition	A single experienced radiologist	1–66 days	BI-RADS 4-category classification: Kappa 0.76 (95% CI: 0.676-0.842); quadratic weighted kappa 0.90 (95% CI: 0.860-0.938). 2-category: 0.815 (0.746, 0.885)
Harvey 2013[49]	BIRADS 3rd edition (prior to 2003) or 4th edition (released in 2003); not shown separately	Radiologist	Mean 429 days (around 14.3 months) +/- 127 days	Linear weighted κ value (95% CI): 0.544 (0.540, 0.549)*; quadratic weighted kappa: 0.638 (0.634, 0.642)* *=calculated by CS
Holland 2016[40]	BIRADS 4th edition and Volpara v 1.5.0	Three radiologists with more than eight years of experience in breast imaging; PhD student with a medical degree and two years of experience with breast imaging. The radiologists were familiar with the density categories, as these are routinely assessed in clinical practice.	30 months	The agreement was substantial for the readers for BIRADS with weighted kappa values ranging from 0.76 to 0.82 using four classes (weighting not stated). Radiologists: 0.76, 0.77, 0.79; student: 0.82. The agreement was substantial for the readers for BIRADS with values ranging from 0.68–0.77 using two classes. Using Volpara VDG the authors obtained a weighted kappa of 0.85 (0.82–0.87). Using VDG the authors obtained a kappa of 0.80 (CI 0.74–0.85) for two classes. The ICC (95% CI) of the scores for the prior and current exams was 0.91 (0.89–0.92), 0.79 (0.75–0.82), 0.77 (0.73–0.81), 0.76 (0.72–0.79), 0.82 (0.79–0.84), and 0.75 (0.71–0.78) for VDG, R1, R2, R3, R4 and RG ('group reading', by assigning the score of a randomly chosen reader) respectively.
Irshad 2016[12]	BIRADS 4th edition	Five fellowship-trained radiologists (breast imagers with 3–17 years of experience)	4 weeks	4th-edition BI-RADS: overall intrareader agreement (quadratic weighted kappa) 0.84

	BIRADS 5th edition		4 weeks	(95% CI, 0.80–0.87); individual intrareader agreements in five readers ranged from 0.78 (95% CI, 0.69–0.88) to 0.92 (95% CI, 0.87–0.97); four readers >0.8 and one <0.8. 5th edition BIRADS: overall intrareader agreement 0.77 (95% CI, 0.73–0.81); individual intrareader agreements in five readers ranged from 0.74 (95% CI, 0.64–0.84) to 0.99 (95% CI, 0.98–1.00); four readers >0.8 and one <0.8.
Llobet 2014,[15] Martinez Gomez 2014[54] and Pollan 2013[55]	DM-Scan	3 highly experienced radiologists in screening mammographies. Raters R1 and R2 had been reading screening mammograms from more than 10 years, with 2 years' experience of full digital mammography in the former case and 6 years of indirect digital mammography in the latter. R3 had been reading mammograms for 34 years, including 2 years of indirect digital mammographs and 6 years of full digital mammograms.	2 months	Test-retest ICC (95 % CI) for semi-automated (DM-Scan) estimation: Reader 1: 0.935 [0.911 0.952]; reader 2: 0.938 [0.915 0.955]; reader 3: 0.900 [0.863 0.926]; mean of the three readers: 0.924 [0.896 0.944]
Sprague 2016[22]	BI-RADS 4th edition	83 radiologists	Median, 1.1 years, IQR 1.0 to 1.2 years	Among women with consecutive mammograms interpreted by the same radiologist (n = 11 042 women), 10.0% had discordant ratings for dense versus nondense status at the 2 examinations; linear weighted kappa 0.760 (0.7507, 0.7695)*, quadratic weighted kappa 0.8338 (0.8172, 0.8504)* * Calculated by CS
van der Waal 2015[13]	BI-RADS 5th edition	Three experienced screening radiologists.	Not stated	The κ_w were 0.82 (95% CI: 0.79–0.86), 0.85 (0.80-0.89) and 0.87 (95% CI: 0.83–0.91) for the three readers on a four-category scale.

Table c: Results: Inter-rater reliability

Study	Intervention	Readers	Outcome reported
Abdolell 2013[37]	Visual percent density assessment	Two senior mammographers, one junior mammographer, one senior resident, and one fellow.	ICC = 0.884 (95% CI 0.854, 0.910)

Alshafeiy 2017[48]	BIRADS 5th edition	Three radiologists; 5–25 years of experience in breast imaging.	For digital 2D mammography, on a four-category scale, weighted kappa (weighting not stated): Reader 1 and 2: 0.56 (0.48–0.63) Reader 1 and 3: 0.59 (0.52–0.66) Reader 2 and 3: 0.68 (0.61–0.74) For digital 2D mammography, on a two-category scale: Reader 1 and 2: 0.67 (0.59–0.75) Reader 1 and 3: 0.67 (0.59–0.75) Reader 2 and 3: 0.82 (0.75–0.89) Interreader agreement for the two-category scale was significantly different between readers 1 and 2 and readers 1 and 3 ($p < 0.001$ for both) but not between readers 2 and 3 ($p = 1.000$).
Ekpo 2016.[36]	BI-RADS 5th edition	Three Royal Australian and New Zealand College of Radiology (RANZCR) certified breast radiologists	Cohen's unweighted Kappa (κ) (95% CI) on a four-grade scale: Reader 1 vs. Majority report: 0.79 (0.74–0.85) Reader 2 vs. Majority report: 0.72 (0.66–0.78) Reader 3 vs. Majority report: 0.65 (0.58–0.73) Reader 1 vs. 2: 0.68 (0.61–0.75) Reader 1 vs. 3: 0.58 (0.50–0.65) Reader 2 vs. 3: 0.38 (0.30–0.46) The average of the reader 1 vs. 2, 1 vs. 3 and 2 vs. 3 kappas: 0.55 (0.47–0.62) Cohen's unweighted Kappa (κ) (95% CI) on a two-grade scale: Reader 1 vs. Majority report: 0.94 (0.92–0.97) Reader 2 vs. Majority report: 0.83 (0.75–0.89) Reader 3 vs. Majority report: 0.87 (0.80–0.93) Reader 1 vs. 2: 0.81 (0.72–0.89) Reader 1 vs. 3: 0.85 (0.78–0.92) Reader 2 vs. 3: 0.70 (0.61–0.78) The average of the reader 1 vs. 2, 1 vs. 3 and 2 vs. 3 kappas: 0.79 (0.70–0.86)
Ekpo 2016.[19]	BI-RADS 4th edition	Five Royal Australian and New Zealand College of Radiology-certified breast radiologists. Number of years certified: R1: 13; R2: 20; R3: 3; R4:	A substantial (0.61 to 0.80) to almost perfect (0.81 to 1.00) agreement was observed between the individual radiologist and the majority report

		20; R5: 19; R6: 35 (mean 18.3). Number of years reading scorning mammograms: 13; 20; 3; 20; 19; 25, respectively (mean 16.7).	on a four-grade scale for BI-RADS from 0.80 (95% CI 0.76 to 0.83) to 0.89 (0.84 to 0.93). There was substantial inter-reader agreement (in pairs) on a four-grade scale from weighted kappa (weighting not stated) of 0.66 (0.62 to 0.71), 0.73 (0.68 to 0.77) and 0.75 (0.70 to 0.81). An almost perfect agreement was observed between the individual radiologist and the majority report on a two-grade scale from 0.82 (0.77 to 0.87) to 0.90 (0.85 to 0.94). There was substantial 0.77 (0.73 to 0.82) to almost perfect 0.89 (0.84 to 0.93) inter-reader agreement on a two-grade scale.
Eng 2014[9] and Busana 2016[53]	Cumulus	Not stated (random sample of 200 women whose images were independently read by a second observer)	The ICC for Cumulus percent density was 0.89, 0.90 and 0.83 for raw, processed and analogue-like images, respectively.
Eom 2017[45]	BIRADS 5th edition	Two were breast-imaging experts with more than five years of experience in reading mammograms, two were general radiologists with fewer years of experience in reading mammograms, and two were medical students without clinical experience in breast imaging. Two medical students were trained to read total of 80 mammogram set comprised of 20 mammograms per each Volpara density categories.	The four-category agreement between the expert and general radiologist was moderate (k=0.67). The two-category agreement between visual assessment of the expert and general radiologist was substantial (k=0.78). BI-RADS density 4-category weighted kappa (weighting not stated) Breast-imaging expert vs. general radiologist 0.67 (0.63 to 0.70) General radiologist vs. student 0.02 (-0.02 to +0.06) Breast-imaging expert vs. student 0.00 (-0.04 to +0.04) Non-dense vs. dense Breast-imaging expert vs. general radiologist 0.78 (0.73 to 0.82) General radiologist vs. student 0.03 (-0.02 to +0.09) Breast-imaging expert vs. student 0.00 (-0.04 to +0.05)
Gweon 2013[42]	BI-RADS 4th edition	Three blinded radiologists who specialize in breast imaging and at the time of the study had 5–10 years of experience in interpreting mammography and 5–8 years of experience in softcopy review of digital mammography	The overall weighted kappa (weighting not stated) of the three radiologists' estimates of BI-RADS density categories showed moderate agreement (κ = 0.48).

			Pairwise estimates of the weighted kappa between two different observers showed moderate to substantial agreement (κ = 0.51–0.64).
Holland 2016[40]	BIRADS 4th edition	Three radiologists with more than eight years of experience in breast imaging; PhD student with a medical degree and two years of experience with breast imaging. The radiologists were familiar with the density categories, as these are routinely assessed in clinical practice.	There was a substantial to almost perfect agreement, with weighted kappa values (weighting not stated) between 0.78 and 0.83 using four categories. The agreement for two categories is between 0.73 and 0.78.
Irshad 2016[12]	BIRADS 4th edition BIRADS 5th edition	Five fellowship-trained radiologists (breast imagers with 3–17 years of experience)	The overall interreader agreement (quadratic weighted kappa) using the fourth-edition BI-RADS criteria was 0.65 (95% CI, 0.61–0.69), whereas the overall interreader agreement using the fifth-edition BI-RADS criteria was 0.57 (95% CI, 0.53–0.61). The difference between the interreader agreements obtained using the old and new BI-RADS criteria was statistically significant (p = 0.006). Fleiss-Cohen (Quadratic) Weighted κ (95% CI) for reader pairs ranged from 0.67 (0.56–0.78) to 0.87 (0.80–0.93) for 4th edition and from 0.61 (0.48–0.74) to 0.90 (0.84–0.95) for 5th edition.
Irshad 2017[51]	BIRADS 4th edition	Five radiologists; all fellowship trained in breast imaging with clinical experience ranging from 3 to 15 years in reading mammograms	There was a statistically excellent agreement in the density distribution pattern between the readers for the BIRADS 4th edition (ICC 0.940, 95% CI 0.754 to 0.996).
Kang 2016[43]	Cumulus (version 4.0)	Two radiologists board certified in breast imaging and one breast surgeon (> 10 years of experience in mammographic reading)	All three readers' percentage density estimates agreed with one another for the interactive thresholding method (CCC 0.86-0.89).
Llobet 2014,[15] Martinez Gomez 2014[54] and Pollan 2013[55]	BIRADS 4th edition DM-Scan	Three highly experienced radiologists in screening mammographies. Raters R1 and R2 had been reading screening mammograms from more than 10 years, with 2 years' experience of full digital mammography in the former case and 6 years of indirect digital mammography in the latter. R3 had been reading mammograms for 34 years, including 2 years of indirect digital mammographs and 6 years of full digital mammograms.	The average quadratic weighted kappa was 0.823 (95% CI: 0.818–0.829) in the BI-RADS scale. Inter-rater ICC with their 95 % confidence intervals for semi-automated (DM-Scan) estimation: Reader 1 vs. Reader 2: 0.922 [0.910, 0.933] Reader 1 vs. Reader 3: 0.928 [0.916, 0.938] Reader 2 vs. Reader 3: 0.916 [0.902, 0.927] Mean: 0.922 [0.909, 0.933]
Lobbes 2012[16]	BIRADS 4th edition	Mammoradiologist: 18 years' experience; senior resident in radiology: 2 years' experience	Inter-rater reliability of experienced versus inexperienced reader: overall linear weighted kappa: 0.521 (95% CI 0.446-0.597); moderate. Quadratic weighted kappa 0.65 (0.53, 0.77)*. * Calculated by CS

			Left versus right breast: CC projection: ICC 0.92, 95% CI 0.89 to 0.94; MLO projection: 0.91, 95% CI 0.89 to 0.93.
Mazor 2016[39]	BIRADS 5th edition	Ten mammography technologists and seven breast radiologists. Technologists: variable levels of experience; seniority, ranging from 12 to 60 months (mean: 29.4 months, SD: 13.2months). Each technologist underwent dedicated training for breast density evaluation according to the 5th edition of the BI-RADS breast density system before participating in the study. Radiologists: at least ten years of experience	Overall, only a fair level of agreement was noted between the technologists and the radiologists in determining BDS, with a weighted kappa (weighting not stated) of 0.38 (95% CI: 0.33, 0.43) using four categories. For four categories: Technologists only: 0.62 (95% CI: 0.53, 0.71) Radiologists only: 0.69 (95% CI: 0.59, 0.78) For two categories: kappa value of 0.45 (95% CI: 0.38, 0.51) between the technologists and the radiologists, indicating a moderate level of agreement. Fewer women were evaluated with breast density scores of 1–2 by the technologists (49%) as compared to the radiologists (73%). Conversely, the technologists evaluated more women with the higher breast density scores of 3–4 (51%) as compared with the radiologists (27%). For two categories: Technologists only: 0.62 (95% CI: 0.49, 0.74) Radiologists only: 0.77 (95% CI: 0.66, 0.87)
Osteras 2016[41] and Osteras 2016[56]	BIRADS 4th edition	Five radiologists: 11, 34, 24, 1 and 3 years' experience (radiologists 1-5 respectively)	BIRADS: Four of the five radiologists had almost perfect agreement with the median score using quadratic weights: Radiologist 1: 0.879 (0.855-0.901) Radiologist 2: 0.875 (0.848-0.900) Radiologist 3: 0.849 (0.823-0.873) Radiologist 4: 0.934 (0.915-0.951) Radiologist 5: 0.763 (0.724-0.798) BIRADS: Using unweighted kappa with four categories, four of five radiologists showed substantial agreement or better, while one showed moderate agreement. Radiologist 1: 0.724 (0.675-0.771) Radiologist 2: 0.748 (0.701-0.794) Radiologist 3: 0.672 (0.619-0.722)

			Radiologist 4: 0.856 (0.817-0.891) Radiologist 5: 0.525 (0.465-0.582)
Sartor 2016[47]	BIRADS 4th edition	Five breast radiologists; all had >10 years' experience in breast radiology.	There was substantial agreement between BIRADS scores with a linear weighted kappa of 0.77 (0.76 to 0.79); percent of observations on which raters agreed 80.9%.
Singh 2016[38]	BIRADS 5th edition	Two blinded radiologists who specialize in breast imaging; 5-10 years of experience in interpreting mammography	BIRADS: almost perfect agreement (κ = 0.895). 444/476 examinations (93.3%) showed agreement between the two observers; the other 32 showed differences within 1 category only.
Sprague 2016[22]	BI-RADS 4th edition	Eighty-three radiologists	Among women with consecutive mammograms interpreted by different radiologists (n = 34 271 women), 32.6% had a different density assessment at the 2 examinations. With density dichotomised as dense or nondense, 17.2% of women with consecutive mammograms interpreted by different radiologists had discordant density ratings at the 2 examinations; 27.0% of women with dense breasts at the first examination were deemed to have nondense breasts at the second examination, and 11.4% of women with nondense breasts at the first examination were deemed to have dense breasts at the second examination. The median percentage of mammograms rated as showing dense breasts (heterogeneously or extremely dense) was 38.7%, with an interquartile range of 28.9% to 50.9% and a full range of 6.3% to 84.5%. Twenty-five percent of radiologists rated fewer than 28.9% of their patients' mammograms as showing dense breasts, whereas the highest 25% of radiologists rated at least 50.9% of their patients' mammograms as showing dense breasts.
van der Waal 2015[13]	BI-RADS 5th edition	Three experienced screening radiologists.	The mean proportion of agreement for the pair-wise comparisons was 71.3% (range %: 67.6–74.3, range n: 671–737). The quadratic κ_w of the inter-rater comparisons ranged from 0.80 to 0.84, which corresponds to 'good' or 'very good' reliability. The mean proportion of agreement for the pair-wise comparisons when the measure was dichotomised was higher (range %: 89.0–90.2).

Table d: Results: Concordance

Study	Intervention/comparator	Readers	Outcome reported
Abdolell 2013[37]	Densitas vs. median of the visual % density assessments performed by the five participating radiologists	Two senior mammographers, one junior mammographer, one senior resident, and one fellow.	ICC = 0.862 Bland-Altman: bias = 1.86% (95% CI not explicitly reported. Says "both were less than 25%"), lower limit of agreement = -20.38, upper limit of agreement = 24.1, largest outlier = not reported
Conant 2017[17]	LIBRA vs. BIRADS 5th edition	Radiologist	There was a correlation between the increasing BIRADS categories and increasing mean percent density estimates using LIBRA; shown graphically.
Destounis 2017[18]	BIRADS 4th edition, from previous normal mammogram vs. Volpara v1.4.2 from previous normal mammogram if raw images available or contralateral breast if raw images not available	Radiologists; breast imaging experience ranged from 6 to 35 years	Linear weighted κ = 0.512 Kappa recalculated for the review (CS) using quadratic weights (κ = 0.652, 95% CI 0.56, 0.744) rather than linear weights (κ = 0.512 95% CI 0.466, 0.557)
Ekpo 2016.[19]	Quantra vs. BI-RADS 4th edition majority report	All Royal Australian and New Zealand College of Radiology-certified breast radiologists. Number of years certified: R1: 13; R2: 20; R3: 3; R4: 20; R5: 19; R6: 35 (mean 18.3). Number of years reading scorning mammograms: 13; 20; 3; 20; 19; 25, respectively (mean 16.7).	Simple kappa four-grade scale: 0.55 (0.48–0.63) Weighted kappa four-grade scale 0.79 (0.75–0.84) Simple kappa two-grade scale: 0.57 (0.50–0.64) Weighted kappa two-grade scale): 0.84 (0.79–0.87)
Eng 2014[9] and Busana 2016[53]	BI-RADS 4th edition; Cumulus v3; ImageJ-based method; Volpara v1.0; Quantra v1.3; single energy x-ray absorptiometry (SXA) method, v6.5	Not stated	Bland-Altman plots showed no systematic differences in square root transformed Cumulus and LIBRA percent density values from the same type of image. In all, 45–47 % of women were assigned to the same quintile and 81–87 % to the same ±1 quintile by LIBRA and Cumulus percent density estimates on the same type of image. Cumulus vs. Quantra: 52% of women assigned to the same quintile Cumulus vs. SXA: 48% assigned to the same quintile Cumulus vs. Volpara: 55% assigned to the same quintile Quantra vs. SXA: 50% assigned to the same quintile Quantra vs. Volpara: 66% assigned to the same quintile

Eom 2017[45]	BIRADS 5th edition vs. Volpara version 1.5.12	Two were breast-imaging experts with more than five years of experience in reading mammograms, two were general radiologists with fewer years of experience in reading mammograms, and two were medical students without clinical experience in breast imaging. Two medical students were trained to read total of 80 mammogram set comprised of 20 mammograms per each Volpara density categories.	The four-category agreement between visual assessments of the breast-imaging expert and volumetric assessments by Volpara was substantial (k=0.77). The agreement between visual assessments by the student and volumetric assessments by Volpara was slight (k=0.01). The two-category agreement between visual assessments of the breast-imaging expert and volumetric assessments by Volpara was almost perfect (k=0.83). The agreement was substantial between visual assessments of general radiologist and volumetric assessment by Volpara (k=0.73), but the agreement between visual assessments of the students and volumetric assessments by Volpara was slight (k=0.01). BIRADS 4-category: Reader vs. Volpara Breast-imaging expert 0.77 (0.75 to 0.80) General radiologist 0.71 (0.68 to 0.74) Student 0.01 (-0.04 to +0.05) Non-dense vs. dense: Reader vs. Volpara Breast-imaging expert 0.83 (0.80 to 0.87) General radiologist 0.73 (0.68 to 0.77) Student 0.01 (-0.05 to +0.07)
Gweon 2013[42]	BIRADS 4th edition; Volpara version 1.5.1	Three blinded radiologists who specialize in breast imaging and at the time of the study had 5–10 years of experience in interpreting mammography and 5–8 years of experience in softcopy review of digital mammography	Pairwise estimates of the weighted kappa between BI-RADS density category by two radiologists' agreement and Volpara VDG showed moderate agreement (κ = 0.54 reported in paper; linear weighted kappa: 0.5276 (0.4824, 0.5728)*; quadratic weighted kappa: 0.6471 (0.5495, 0.7447)*). *=calculated by CS

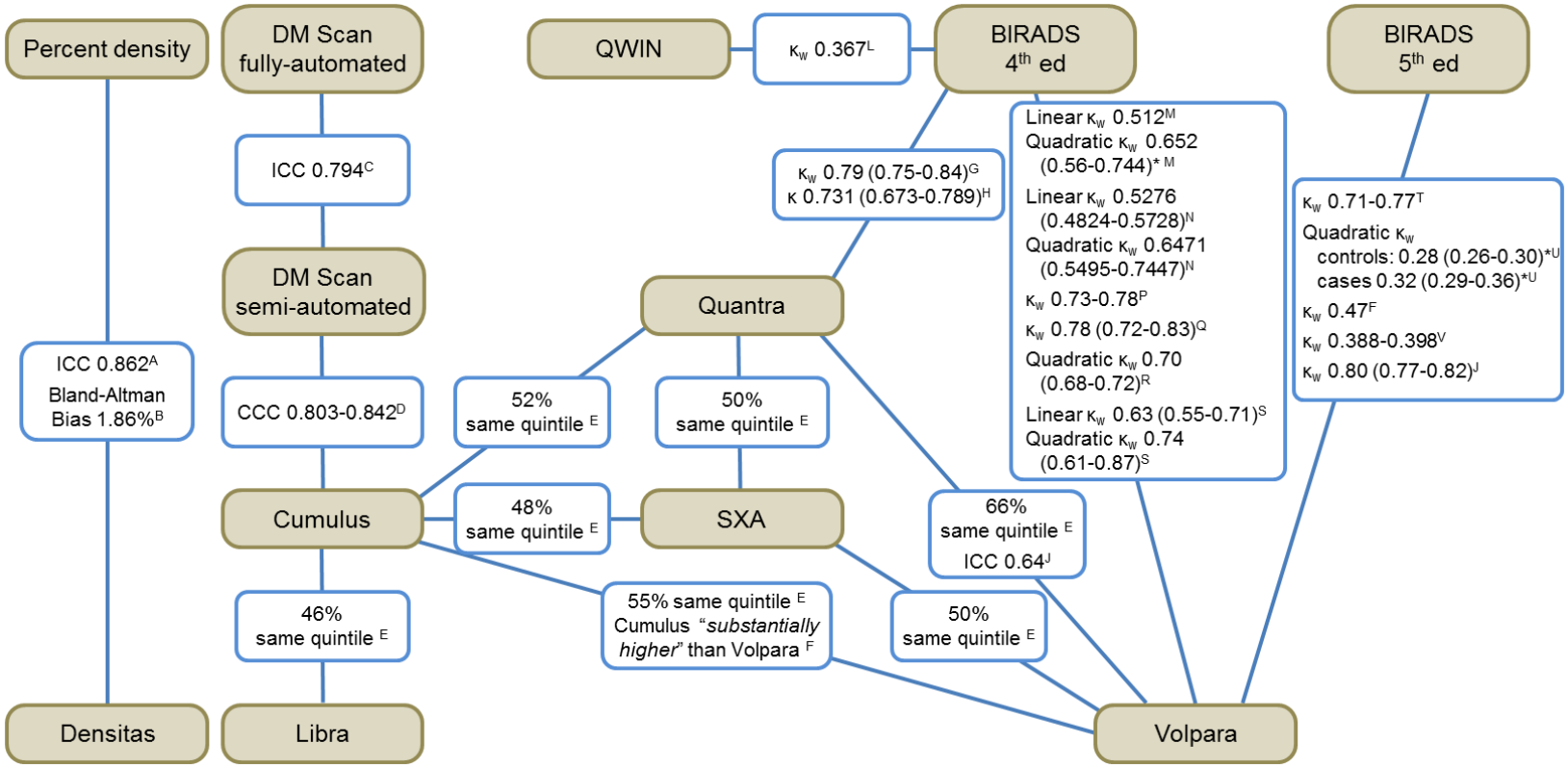
Holland 2016[40]	BIRADS 4th edition and Volpara v 1.5.0	Three radiologists with more than eight years of experience in breast imaging; PhD student with a medical degree and two years of experience with breast imaging. The radiologists were familiar with the density categories, as these are routinely assessed in clinical practice.	The agreement between the readers and VDG is lower than the inter-reader agreement; with kappa values between 0.73 and 0.78 using four categories. In most of the pairs with a disagreement between VDG and the reader, a higher score was given by the software than by the reader. The agreement between the readers and VDG is lower than the inter-reader agreement; with kappa values between 0.63 and 0.71 using two categories.
Jeffers 2017[14]	Cumulus 6 (version 4.0); Volpara (version not stated) and BI-RADS (version not stated)	A single reader (with 2 years of experience), who was blinded to whether the images were for patients or control subjects, performed all Cumulus measurements. The reader was trained by the providers of the Cumulus software. Readers for Volpara and BI-RADS not stated.	The agreement of clinical BIRADS and Volpara density categorisations was fair, with a weighted kappa statistic of 0.47. Cumulus area-based percentage of density measurements were substantially higher than were Volpara volumetric percentage of density measurements.
Kerlikowske 2017[52]	BIRADS 5th edition; Volpara version 1.5.0	Practising radiologists	A wide distribution of dense breast volume was observed within each BI-RADS density category. Surprisingly, about one-third (30.5%) of control subjects with almost entirely fat breasts had first-quartile dense breast volume (≤35.9 ml), and about half (54.1%) with extremely dense breasts had fourth-quartile (>70.0 ml) dense breast volume. The correlation coefficient between continuous dense breast volume and BI-RADS density was r = 0.38 (95% CI 0.34–0.42) for cases and r = 0.31 (95% CI 0.29–0.34) for control subjects. Weighted (quadratic) kappa = 0.28 (0.26, 0.30)* for control subjects; weighted (quadratic) kappa = 0.32 (0.29, 0.36)* for case subjects. *=calculated by CS
Llobet 2014,[15] Martinez Gomez 2014[54] and Pollan 2013[55]	DM-Scan semi-automated vs. DM-Scan fully automated	3 highly experienced radiologists in screening mammographies. Raters R1 and R2 had been reading screening mammograms from more than 10 years, with 2 years' experience of full digital mammography in the former case and 6 years of indirect digital mammography in the	ICC (95% CI) comparing the fully-automated and the semi-automated (DM-Scan) methods for each rater: Reader 1: 0.800 [0.771, 0.826] Reader 2: 0.838 [0.814, 0.860] Reader 3: 0.785 [0.754, 0.813]

	Cumulus vs. DM-Scan	latter. R3 had been reading mammograms for 34 years, including 2 years of indirect digital mammographs and 6 years of full digital mammograms.	Mean: 0.794 [0.764, 0.821] Concordance Correlation Coefficient (CCC) (95% CI): Reader 1: 0.841 (0.820 to 0.863) Reader 2: 0.803 (0.777 to 0.828) Reader 3: 0.842 (0.820 to 0.864)
Lobbes 2012[16]	BIRADS 4th edition vs. QWIN	Mammoradiologist: 18 years' experience; senior resident in radiology: 2 years' experience	Experienced reader: $\kappa = 0.367$
Osteras 2016[41] and Osteras 2016[56]	Quantra vs. BIRADS 4th edition	5 radiologists: 11, 34, 24, 1 and 3 years' experience (radiologists 1-5 respectively)	Quantra (at 10% threshold) versus radiologists median BIRADS 4th edition: Binary classification (unweighted) kappa = 0.731 (0.673-0.789) 12 (2.2%) were unanimously scored fatty by radiologists and dense by Quantra (false positives); 2 (0.4%) were unanimously scored dense by radiologists and fatty by Quantra (false negatives).
Raza 2016[50]	Agreement between the "Leads truth" (LT) from breast imagers using BIRADS 4th editions vs. Volpara ("quantitative truth" [QT])	Two senior breast imagers, each with more than 20 years of breast imaging experience	The quantitative density tool tended to assign higher density categories to the 200 cases than the study leads assigned. The calculated weighted k statistic was 0.78 (95% CI, 0.72 to 0.83) indicating substantial agreement.
Sartor 2016[47]	BIRADS 4th edition; Volpara (version 1.5.11)	5 breast radiologists; all had >10 years' experience in breast radiology	Agreement between Volpara density grade (VDG) and BIRADS per radiologist: linear weighted kappa: Radiologist 1: 0.66 (0.56, 0.75) Radiologist 2: 0.56 (0.54, 0.58) Radiologist 3: 0.48 (0.44, 0.52) Radiologist 4: 0.52 (0.48, 0.56) Radiologist 5: 0.57 (0.53, 0.61) Overall: 0.55 (0.53, 0.56) Overall quadratic weighted kappa: 0.7004 (0.6842, 0.7166)* * Calculated by CS
Seo 2013[44]	BIRADS 4th edition and Volpara (version 1.4)	Two board certified radiologists who each had several years of experience in reading mammograms (17 years and 7 years) and a 3rd-year radiology resident	There were 134 cases of agreement and 59 cases of disagreement (30.6%; 54 were over-scored using VDG and 5 under-scored).

			Linear weighted kappa = 0.63 (0.55, 0.71)* Quadratic weighted kappa = 0.74 (0.61, 0.87)* * Calculated by CS
Singh 2016[38]	BIRADS 5th edition and Volpara (version 1.4.5)	2 blinded radiologists who specialize in breast imaging; 5-10 years of experience in interpreting mammography	Pairwise estimates of weighted kappa between VDG grade and BIRADS density by 2 observers showed fair agreement (κ = 0.398 and 0.388, respectively). On visual assessment, <25% of the study population was categorised as BIRADS 3 or 4, whereas Volpara assigned around 41% to the dense category.
van der Waal 2015[13]	Volpara (version 1.5.11) vs. BI-RADS 5th edition Volpara (version 1.5.11) vs. Quantra (version 1.3)	Three experienced screening radiologists.	The Volpara VDG distribution was comparable to the BI-RADS density distribution (κ_w: 0.80, 95% CI: 0.77–0.82; proportion agreement: 65.4%). The median volumetric percent density was 12.1% (IQR: 9.6–16.5) for Quantra, which was higher than the Volpara estimate (median 6.6%, IQR: 4.4–10.9). The mean difference between Quantra and Volpara was 5.19% (95% CI: 5.04–5.34) (ICC: 0.64).

Figure e: Diagram of concordance (excluding untrained students)

While a Kappa of 1 represents a perfect agreement, Kappa values of 0 or below represent agreements that occur by chance, or that are poor. Kappa values of 0.01–0.20 represent slight agreement, values of 0.21–0.40 represent fair agreement, those between 0.41–0.60 represent moderate agreement, values between 0.61–0.80 represent substantial agreement and values between 0.81–0.99 represent almost perfect agreement. ICC is equivalent to weighted kappa. Concordance between methods was fair to substantial.



* Kappa calculated

CCC = Concordance Correlation Coefficient ICC = Intraclass correlation coefficient; κ = Unweighted Kappa; κ_w = Weighted Kappa

A: Abdolell 2013
B: [*limits of agreement -20.38 to +24.1, largest outlier not reported*]
C: Llobet 2014
D: Pollan 2013
E: Eng 2014 & Busana 2016
F: Jeffers 2017
G: Ekpo 2016 [*Quantra…*]
H: Osteras 2016 [*Classification...*]
J: van der Waal 2015
L: Lobbes 2012. [*Experienced reader*]
M: Destounis 2017
N: Gweon 2013
P: Holland 2016
Q: Raza 2016
R: Sartor 2016
S: Seo 2013
T: Eom 2017
U: Kerlikowske 2017
V: Singh 2016

Question 2a

Table a: Design and limitations

Study	Population (n)	Interventions/ Comparator	Outcome	No. centres; country	Limitations
Destounis 2017[18]	Women aged >40 years (mean 62.1; SD 11) with histopathologically confirmed breast cancer (n=614)	Mammographic density using BIRADS 4th edition or Volpara	Comparison between screen-detected and interval cancers	1; USA	Retrospective study; BMI not available and so not included in multivariate analysis. Interval cancers not differentiated between true interval, missed or mammographically occult (i.e. masked by dense tissue). Unable to analyse the relation between masking risk and location and distribution of density within the breast. Large proportion of people missing from analysis. Around 13.6% aged <50 years and 23.6% >70 years. Around 8.5% <47 years and 16.1% >73 years.
Holland 2017[61]	Cases: Women with interval cancers within 12 months after the examination. The last available screening examination before cancer diagnosis is used in this study. Mean age 57.7 years. Controls: For each patient with an interval cancer, 10 participants were chosen as controls. The control participants needed to have had a mammographic examination in the same month in which the last screening examination of the interval cancer patient was performed. To be eligible as control, the women should not have been recalled on the basis of this mammographic examination and they should not have been diagnosed with breast cancer within 2 years after this	Percent dense volume using Volpara or percent density using BIRADS 5th edition	To measure to what extent the methods can identify women at high masking risk, the mammograms were divided in a high and low masking risk group by thresholding the risk measure. Then, the sensitivity of the masking measures was computed as the number of interval cancers in the high-risk group divided by the total number of interval cancers. The false positive rate is calculated as the percentage of normal controls selected as at high masking risk at the same threshold. In the context of risk stratification for supplemental screening, the proportion of controls selected as at high masking risk can be seen as supplemental screening rate and the	1; The Netherlands	Given that the exact cancer location was unknown and that the diagnostic mammograms were not available, it was not possible to review the interval cancers and to confirm that masking is the cause for a cancer diagnosis outside the screening program. CC images were not available for all exams. BIRADS density assessments of only one radiologist were available. Many studies found inter- and intra-reader variability in breast density assessment using BI-RADS. Therefore, to make a definitive comparison between the automated methods and radiologists assessments, an extensive reader study should be conducted with multiple readers.

	examination. Controls without a density map, due to failure of the computation, were replaced. (n=111 cases + 1110 controls). Mean age 59.2 years.		proportion of interval cancers gives an estimate about the cancers that might be detectable with additional imaging at that supplemental screening rate.		
Kerlikowske 2015[62]	Women aged 40-74 years who did not have a history of breast cancer or breast implants and had complete information on demographic and breast health history information (n=365,426)	Mammographic density using BIRADS	Interval cancer rate and false positive rate by breast density	Not stated; USA	The cut-points used for defining low performance were developed for identifying minimally acceptable performance levels for screening mammography interpretation for invasive and DCIS outcomes combined; the authors state that they do not know if these performance cut-points are related to long-term outcomes such as breast cancer mortality. For some subgroups with an average interval cancer rate <1/1,000 mammograms, they cannot rule out a higher interval cancer rate because the upper 95% confidence limit exceeds one. A 24-month interval was not evaluated since women may return early for screening and/or have mammograms outside the BCSC. Participation rate not stated. 19.1% aged 40-49 years and 13.4% aged 70-74 years
Nelson 2016[59]	Women aged 40 to 89 years who had routine screening with digital mammography (n=405,191)	Mammographic density using BIRADS 4th edition	Rates of false-positive and false-negative mammography results and recommendations for additional imaging and biopsies from a single screening round	5 registries; USA	The BCSC data reflect opportunistic screening in a fluctuating population of women in the U.S. whose information was collected by the participating registries. Findings may not be applicable to other populations. Limitations also include restrictions of registry data with pre-defined data elements and the inherent biases of observational data. Some outcomes, such as the effectiveness and

					harms of different screening intervals, would be more accurately determined by comparing outcomes between women who were randomly assigned to comparison groups. 16.3% had missing data for breast density. 28.1% aged 40–49 years, 12.4% aged 70–79 years and 4.6% aged 80–89 years.
Rawashdeh 2013[58]	A single-image bank containing 60 digital cases containing 20 positive (biopsy-proven) cases with a single focus of cancer in 16 cases and multicentric cancer in 4 cases (resulting in a total of 24 cancers) (n=60). Mean 54 years (range 47 to 78 years)	BIRADS 3rd edition	Detectability of lesions by breast density in a reader study	Not stated; Australia	The same radiologist who chose the images was responsible for assessing breast density; <100 images
Timmermans 2017[60]	Women aged between 50 and 69 years (n=351,532)	BI-RADS 4th edition	Cancer detection rate, interval cancer rate, third readings and correlated false-positives by breast density category	Not stated; Belgium	Subdivision of ICs in true, missed and minimal signs was not performed in the present study. A low statistical power hampered reaching statistical significance in differences between modalities for the BI-RADS IV class data.
Wanders 2017[7]	Women aged 50–75 years participating in a biennial screening program (n=111,898 examinations belonging to 53,239 women)	Volpara	Interval cancers by density	1; The Netherlands	A limitation of this study is that during the study period, the MLO view was the standardly acquired view for the subsequent screening rounds and CC views were only taken in addition to MLO during the first screening round or by indication during subsequent rounds. As a result, breast density was determined based on only MLO views for some examinations and on both MLO and CC views for other examinations in our main analysis. Volpara's volumetric percent density measured on CC views tends to be

					somewhat higher than on MLO views. As CC views are more often performed among women with dense breasts and women with a suspicious region on their MLO view, breast density might be somewhat artificially elevated for these women. Our sensitivity analysis using VDG categories based on volumetric percent density from the MLO views only did not lead to different conclusions. Screening sensitivity is presumably higher when both MLO and CC views are available compared to MLO views only. Therefore, standardly taking both MLO and CC views would lead to higher sensitivity, particularly in women with fatty breasts as they are the ones who most often receive MLO views only. This would lead to larger differences in screening performance across breast density categories.

Table b: Mammographic sensitivity and risk of interval cancers by density

Study	Mammographic sensitivity by density	Risk of interval cancers by density			Risk of bias
		Unadjusted	**Age-adjusted**	**Adjusted for risk factors apart from age**	
Destounis 2017[18]	Mammographic sensitivity by BIRADS density: Fatty replaced 82% Scattered fibroglandular 90% Heterogeneously dense 84% Extremely dense 66% R^2 = 0.463 Mammographic sensitivity by automated density:	In univariate analysis, density was associated with the risk of diagnosis of interval cancer versus screen-detected cancer. BIRADS 3 vs. 1 or 2: OR 1.91 (1.07-3.40), p=0.028 BIRADS 4 vs. 1 or 2: OR 5.00 (2.43-10.33), p<0.001 Volpara automated density grade 3 vs. 1 or 2: OR 1.94 (1.10-3.43), p=0.021	BIRADS 3 vs. 1 or 2: OR 1.60 (0.89-2.89) BIRADS 4 vs. 1 or 2:	After adjustment for age and menopausal status, density was the only risk factor	High: 20% women excluded for unclear reasons

	Grade 1 95% Grade 2 89% Grade 3 83% Grade 4 65% R^2 = 0.914	Volpara automated density grade 4 vs. 1 or 2: OR 5.60 (2.99-10.47), p<0.001 Volpara volumetric breast density quartile 2 vs. quartile 1: OR 1.73 (0.72-4.13) Volpara volumetric breast density quartile 3 vs. quartile 1: OR 2.08 (0.90-4.83) Volpara volumetric breast density quartile 4 vs. quartile 1: OR 5.58 (2.61-11.93), p<0.001	OR 3.82 (1.82-8.06), p<0.001 Volpara automated density grade 3 vs. 1 or 2: OR 1.64 (0.92-2.94) Volpara automated density grade 4 vs. 1 or 2: OR 4.14 (2.13-8.03), p<0.001 Volpara volumetric breast density quartile 2 vs. quartile 1: OR 1.67 (0.70-4.01) Volpara volumetric breast density quartile 3 vs. quartile 1: OR 1.85 (0.79-4.33)	significantly associated with interval cancer rather than screen-detected cancer. BIRADS 3 vs. 1 or 2: OR 1.58 (0.87-2.86) BIRADS 4 vs. 1 or 2: OR 3.60 (1.69-7.69), p<0.001 Volpara automated density grade 3 vs. 1 or 2: OR 1.66 (0.92-2.98) Volpara automated density grade 4 vs. 1 or 2: OR 3.90 (1.99-7.64), p<0.001 Volpara volumetric breast density quartile 2 vs. quartile 1: OR	

			Volpara volumetric breast density quartile 4 vs. quartile 1: OR 4.17 (1.89-9.21), p<0.001	1.62 (0.67-3.88) Volpara volumetric breast density quartile 3 vs. quartile 1: OR 1.85 (0.79-4.35) Volpara volumetric breast density quartile 4 vs. quartile 1: OR 3.96 (1.79-8.80), p=0.001	
Holland 2017[61]	-	With BI-RADS, 427/1110 = 38.5% (CI 35.7–41.3) of the controls (no cancer) had dense breasts. Of the women developing interval cancers, 70/111 = 63.0% (CI 53.5–72.0) were classified as dense. RR of dense breasts among those with interval cancer = 63/38.5 = 1.64 Cannot calculate OR of cancer in dense vs. non-dense breasts as this was a case-control study so proportions of cancers/non-cancers were selected, not the proportions that would occur in a population.	-	-	Moderate: little information on confounders
Kerlikowske 2015[62]	-	No invasive cancer N (%) / Invasive interval cancer within 12 months of screening mammography N (%) Almost entirely fat: 96,608 (11.7) / 214 (7.9) Scattered fibroglandular densities: 338,882 (40.9) / 1084 (40.2)	-	-	Moderate: unclear how many women excluded; little information on confounders. Not generalisable to our population as

		Heterogeneously dense 326,568 (39.4) 1178 (43.7) Extremely dense 66,701 (8.0) 220 (8.2) Odds of cancer in dense breasts = (1178+220)/(326568+66701) = 0.00355 Odds of cancer in non-dense breasts = (214+1084)/(96608+338882) = 0.00298 Odds ratio of cancer in dense vs. non-dense breasts = 0.00355/0.00298 = 1.19 Interval cancer rate per 1000 mammograms (95% CI). Bold numbers outside minimally accepted cut-points: interval cancer rate >1/1000 mammograms **BI-RADS breast density** **Age (years)** / **Almost entirely fat** / **Scattered fibroglandular densities** / **Heterogeneously dense** / **Extremely dense** 40 – 49: 0.19 (0.04, 0.56) / 0.26 (0.16, 0.40) / 0.76 (0.61, 0.93) / 0.98 (0.67, 1.37) 50 – 59: 0.14 (0.05, 0.34) / 0.33 (0.23, 0.45) / 0.80 (0.65, 0.98) / **1.11** (0.72, 1.64) 60 – 69: 0.23 (0.10, 0.45) / 0.49 (0.37, 0.65) / 0.96 (0.75, 1.22) / **1.13** (0.54, 2.09) 70 – 74: 0.35 (0.10, 0.90) / 0.55 (0.33, 0.86) / **1.15** (0.73, 1.72) / **3.45** (1.27, 7.50) Rate goes up by density at all ages.			19.1% aged 40-49 years and 13.4% aged 70-74 years
Nelson 2016[59]	Women with almost entirely fat and scattered fibroglandular densities had lower rates of false-negative mammography results than those with other types of breast density for ages 40 to 69 years. Rates of false-negative digital mammography results by different ways of dividing up the breast density categories	-	-	-	Moderate: number excluded not stated; age, BMI, ethnicity and menopausal status

	(Number per 1,000 women screened per round and 95% CI; option C is the BIRADS categorisation): 40-49 p Women screened, n 113,770 A Fat-Scattered 0.4 (0.3, 0.6) <0.001 Heterogeneous 1.3 (1.0, 1.7) Extreme 1.7 (1.2, 2.5) B Fat 0.2 (0.0, 0.9) <0.001 Scattered 0.5 (0.3, 0.7) Heterogeneous-Extreme 1.4 (1.2, 1.8) C Fat 0.2 (0.0, 0.9) <0.001 Scattered 0.5 (0.3, 0.7) Heterogeneous 1.3 (1.0, 1.7) Extreme 1.7 (1.2, 2.5) D Fat-Scattered 0.4 (0.3, 0.6) <0.001 Heterogeneous-Extreme 1.4 (1.2, 1.8) 50-59 years p Women screened, n 127,958 A Fat-Scattered 0.6 (0.4, 0.8) 0.002 Heterogeneous 1.4 (1.0, 2.0) Extreme 1.6 (0.9, 2.8) B Fat 0.3 (0.1, 0.7) <0.001 Scattered 0.7 (0.5, 0.9) Heterogeneous-Extreme 1.5 (1.1, 1.9) C Fat 0.3 (0.1, 0.7) <0.001 Scattered 0.7 (0.5, 0.9) Heterogeneous 1.4 (1.0, 2.0) Extreme 1.6 (0.9, 2.8) D Fat-Scattered 0.6 (0.4, 0.8) <0.001 Heterogeneous-Extreme 1.5 (1.1, 1.9) 60-69 years p Women screened, n 94,507 A Fat-Scattered 0.8 (0.5, 1.1) 0.006 Heterogeneous 1.7 (1.3, 2.3) Extreme 1.2 (0.6, 2.7) B Fat 0.6 (0.2, 1.5) 0.007 Scattered 0.8 (0.6, 1.2) Heterogeneous-Extreme 1.6 (1.2, 2.2)				measured but only age reported. 16.3% had missing data for breast density. 28.1% aged 40–49 years, 12.4% aged 70–79 years and 4.6% aged 80–89 years.

	C Fat	0.6 (0.2, 1.5)	0.02			
	Scattered	0.8 (0.6, 1.2)				
	Heterogeneous	1.7(1.3, 2.3)				
	Extreme	1.2 (0.6, 2.7)				
	D Fat-Scattered	0.8 (0.5, 1.1)	0.002			
	Heterogeneous-Extreme	1.6 (1.2, 2.2)				
		70-79 years	p			
	Women screened, n	50,204				
	A Fat-Scattered	1.0 (0.6, 1.5)	0.01			
	Heterogeneous	2.3 (1.6, 3.4)				
	Extreme	5.6 (2.4, 12.9)				
	B Fat	0.3 (0.1, 1.1)	0.001			
	Scattered	1.2 (0.7, 1.9)				
	Heterogeneous-Extreme	2.6 (1.8, 3.7)				
	C Fat	0.3 (0.1, 1.1)	0.002			
	Scattered	1.2 (0.7, 1.9)				
	Heterogeneous	2.3 (1.6, 3.4)				
	Extreme	5.6 (2.4, 12.9)				
	D Fat-Scattered	1.0 (0.6, 1.5)	0.003			
	Heterogeneous-Extreme	2.6 (1.8, 3.7)				
		80-89 years	p			
	Women screened, n	18,752				
	A Fat-Scattered	0.9 (0.5, 1.6)	0.25			
	Heterogeneous	1.1 (0.5, 2.4)				
	Extreme	6.9 (2.5, 18.5)				
	B Fat	0.4 (0.1, 3.1)	0.14			
	Scattered	1.0 (0.6, 1.7)				
	Heterogeneous-Extreme	1.7 (0.8, 3.3)				
	C Fat	0.4 (0.1, 3.1)	0.17			
	Scattered	1.0 (0.6, 1.7)				
	Heterogeneous	1.1 (0.5, 2.4)				
	Extreme	6.9 (2.5, 18.5)				
	D Fat-Scattered	0.9 (0.5, 1.6)	0.18			
	Heterogeneous-Extreme	1.7 (0.8, 3.3)				
Rawashdeh 2013[58]	There was a negative correlation between lesion detection on mammography and breast density (r = -0.64, P = .007)			-	-	-	High: selected images in a reader study; age reported

					but no other details
Timmermans 2017[60]	-	There is a systematic increase of interval cancer rate with breast-density class. The percentage of cancers detected in the screening programme over the total number of cancers registered decreases from 84% for density class I to 46% for class IV.	-	-	Moderate: Age range of screening programme stated but no details of sample in terms of mean age, BMI, ethnicity or menopausal status
Wanders 2017[7]	Sensitivity of screening (%): VDG 1: 85.7% (78.1; 91.0) VDG 2: 77.6% (73.2; 81.5) VDG 3: 69.5% (64.1; 74.4) VDG 4: 61.0% (51.2; 70.0) P<0.001	Interval breast cancer rates were higher in higher breast density categories compared to lower density categories with a significant linear trend (p-trend<0.001). Interval cancer rates in the first year after a screening examination were 0.2, 0.8, 1.2, and 2.9% (p-trend<0.001) in Volpara Density Grade (VDG) categories 1, 2, 3, and 4, respectively. All years: Interval cancer/1000: VDG1: 0.7 (0.4; 1.1); VDG 2: 1.9 (1.5; 2.3); VDG 3: 2.9 (2.3; 3.5); VDG 4: 4.4 (3.2; 6.0); p<0.001	-	-	Moderate: No information on BMI, ethnicity or menopausal status

Question 2b

Table a: The identified systematic reviews and the extent to which their methods matched the scope of our review.

	Our scope:	Bae 2016[65]	Huo 2014[66]	Elias 2014[67]	Antoni 2013[68]	Cummings 2009[64] and McCormack 2006[69]
Question	Q2b: **Is mammographic breast density a risk factor for developing breast cancer?**	This meta-analysis investigated the association between breast density in mammography and breast cancer risk in Asian women.	To critically review the current literature on mammographic density (MD) and summarize the current evidence for its association with breast cancer (BC).	Features (including density) related to HER2 overexpression (a marker of cancer aggressiveness)	A systematic review of studies of mammographic density (MD) in relation to risk of subtype-specific breast cancer, by ER, PR, and HER2 status or gene expression profiles.	To review prospective studies about models and sex hormone levels to assess breast cancer risk and use meta-analysis with random effects models to summarize the predictive accuracy of breast density.
Population	Women aged 50-70 attending breast cancer screening from the general population (not specifically chosen high-risk groups) with a population prevalence similar to the UK	Asian women. Seven datasets were of premenopausal women and eight were of postmenopausal women	Not stated	Not stated	Age range in included studies 40-84 years	Not reported
Density measurements	BI-RADS scale scored by a single qualified reader BI-RADS scale scored by a group consensus of readers • Volpara • Quantra • Cumulus • ImageJ • Single energy x-ray absorptiometry (SXA) • DM-Density M-Vu Breast Density • Absolute fat volume	Wolfe classification; percent density (%); DA, density area (cm^2); MDA, mean dense area (cm^2); TBA, total breast area (cm^2); VDG, volumetric density grade (%); ADA, absolute dense area (cm^2).	BIRADS, Cumulus, Boyd semi-quantitative scale, computer-assisted method (CAM), Tabar, DM-Scan, automated volumetric breast density, automated measure, percent density, semi-automated technique: threshold technique	BI-RADS	BIRADS, percent density, visual (fatty, mixed/dense), Wolfe or Cumulus in different included studies	One study assessed breast density by use of BI-RADS ratings and four measured percent density, in addition to the studies included in McCormack 2006[69]

	• Absolute fibroglandular volume • Density calculated on a single mammogram view (e.g. MLO) • Density calculated from 2 views (e.g. MLO plus CC) • Others?		(TT), fully automated method (FAM), semi-automated method (SAM), standard mammogram form (SMF)			
Outcomes	Head to head studies (2 or more types of density measurement) • Positive and negative concordance between pairs of tests (presented as 2x2 or YxY tables) • comparison of characteristics of discordant cases: in particular comparison of risk of breast cancer (i.e. do cases measured high risk by Volpara and low risk by quantra have a higher risk of breast cancer than cases measured low risk by volpara and high risk by quantra) and measures of missing cancers at screening such as interval cancers. Single or head to head studies (1 or more types of test) • Proportion of women who have an interval	Effect size based on adjusted odds ratios (adjustment factors not stated)	Mammographic density as a risk factor for breast cancer; association of mammographic density with breast cancer subtypes and tumour characteristics.	Odds ratio of HER overexpression by density categories	Relative risk estimates and their 95% CIs of subtype-specific breast cancer were estimated by individual studies as odds ratios in case–control and case-only studies and as hazard/rate ratios in cohort studies. The most fully adjusted RRs reported were included. Controlling for age was included in eligibility criteria. In case-only studies, we extracted estimates of the ratios of relative risks (RRR) of ER+ versus ER- breast cancer associated with MD categories; if ER+ subtypes were used as the reference group, the inverse of the RRRs and its confidence limits were taken.	Relative risk of breast cancer; all adjusted for age; some studies adjusted for additional factors which were not stated except to say that studies that further adjust for body mass index or weight observed somewhat stronger associations

	cancer after screening by density for each test • Proportion of women who have breast cancer by density for each test (includes reporting of absolute risk which is of particular interest in low density groups) • Distribution of cancer type by risk group for each test • Odds or risk ratios from <u>unadjusted</u> univariable models of density as a predictor of risk • Odds or risk ratios from adjusted multivariate models of density as a predictor of risk • Predictive accuracy of multivariate models including density as a predictor of risk (if time permits).					
Study design	Head to head or single arm studies	Cohort or case control studies	Not stated	Not stated	(i) Case–control/ case-cohort/ cohort studies in which MD in cases, defined by subtype, is compared to non-cases and (ii) case-only designs where age-adjusted MD in ER+ cases is compared to that in ER- cases.	Prospective studies

Limits (language and date)	English; from 2000	Language not stated: up to December 31, 2015	English; date not stated	Stated to be no restrictions (assume this means none for language); date to February 8, 2013	English; 5th June 2012	Language not stated; January 1, 2004, through January 1, 2008
Limitations		Overall ES from all 6 articles not calculated, because the number of articles related to Asian women was small and because the breast density index varied across articles. The subgroup analysis could not include results that were not divided by menopausal status. The analysis of premenopausal women was insufficient for dose-response meta-regression (DRMR). The subjects included only women who were born and lived in Asia (women born in Asia but emigrated overseas excluded). In the case-control studies, the most recent mammogram before breast cancer diagnosis were used, but this does not reflect the fact that breast density changes with age.	Very little information on systematic review methods	The authors did not formally use a quality assessment tool; the results from this meta-analysis reflect univariable associations only, as individual studies did not adjust their results for potential confounders, such as lesion size or histologic breast cancer subtype, thus precluding solid causal inference.	Differences in density assessment methods. Restricted to English-language publications and only found studies conducted in North America and Europe, in predominantly Caucasian women, thus other countries and ethnic groups, particularly at lower breast cancer risk are not included. Additionally, there was the lack of power to analyse combinations of ER and PR status.	The studies reviewed had various designs, populations, and methods of analysing data. Although breast density is a strong risk factor for breast cancer, BI-RADS has only modest reproducibility and more reproducible quantitative approaches are not validated or feasible for clinical use; so increased predictive accuracy may not be applicable to current clinical practice.

Table b: Quality assessment of systematic reviews using AMSTAR criteria

AMSTAR Checklist	Bae 2016[65]	Huo 2014[66]	Elias 2014[67]	Antoni 2013[68]	Cummings 2009[64] and McCormack 2006[69]
1. Was an 'a priori' design provided?	Search strategy etc presented; assume a priori design. Article selection was conducted in accordance with the preferred reporting items proposed for systematic reviews and meta-analyses	No	Not stated	Search strategy etc presented; assume a priori design.	Search strategy etc presented; assume a priori design.
2. Was there duplicate study selection/ data extraction?	Not stated	Not stated	Yes for both selection and data extraction	Yes for data extraction: The RRs for each MD category were extracted independently by two of us (SA and VM). Not stated for study selection	Not stated
3. Was a comprehensive literature search performed?	PubMed and Scopus: the following search formula was applied: [(breast) OR (mammary)] AND [(cancer) OR (neoplasm)] AND [(density) OR (index)] AND [(Asia) OR (women)].	Keywords 'mammographic dens*', 'dense mammary tissue' or 'percent dens*' were used to search the existing literature in English on PubMed and Medline.	We performed a comprehensive systematic literature search of MEDLINE and EMBASE on February 8, 2013 using synonyms for HER2 and the imaging modalities of interest in combination with breast	Medline only. The search criteria aimed to identify publications that contained all three of (i) breast cancer, (ii) mammographic density, and (iii) an indication that subtypes were analyzed; where the following terms related to (i) breast cancer: ''breast cancer'', ''breast neoplasm'', ''breast tumor'', (ii) mammographic density: ''breast density'', ''mammograph* density'',	The systematic review and meta-analysis by McCormack et al. analyzed studies about the association between breast density and risk of breast cancer that were published up to November 30, 2005. To update that review, we surveyed MEDLINE and EMBASE databases from January 1, 2004, through January 1, 2008, by use of the terms "breast density" or "mammographic density" that were cross-

			cancer. The search was without restrictions.	‘‘mammograph* pattern’’, ‘‘parenchymal pattern’’, ‘‘Wolfe’’, ‘‘BIRADS’’ or ‘‘Tabar’’, and (iii) subtypes: ‘‘receptor’’, ‘‘luminal’’, ‘‘basal’’, ‘‘triple negative’’, ‘‘Sorlie’’, ‘‘HER-2’’, ‘‘HER2’’. Studies identified using this search were scrutinised to find out whether (i) they examined the association of interest and (ii) age had been controlled for either through design features (via matching on age or restricting to a narrow age range) or through adjustment.	referenced with the MeSH term “breast neoplasm” and the free text term “breast cancer.”
4. Was the status of publication (i.e. grey literature) used as an inclusion criterion?	No grey literature	Not stated	No grey literature	Not stated	Not stated
5. Was a list of studies (included and excluded) provided?	Include: yes Excluded: No	No	Include: yes Excluded: No	Include: yes Excluded: No	No
6. Were the characteristics of the included studies provided?	Yes	Yes	Yes	Yes: Tables 1, 2 and 3	No
7. Was the scientific quality of the included studies	No	No	No	No	No

assessed and documented?					
8. Was the scientific quality of the included studies used appropriately in formulating conclusions?	No	No	No	No	No
9. Were the methods used to combine the findings of studies appropriate?	Yes; meta-analysis with consideration of heterogeneity	Narrative only	Yes; meta-analysis with consideration of heterogeneity	No unadjusted meta-analyses; individual studies with age adjustment shown	No unadjusted analyses; all adjusted for age; some studies adjusted for additional factors which were not stated except to say that studies that further adjustment for body mass index or weight observed somewhat stronger associations
10. Was the likelihood of publication bias assessed?	No	No	Yes: Visual inspection of funnel plot asymmetry in combination with Egger tests generally led to a low suspicion for publication bias, albeit the number of studies was sometimes too low for proper evaluation (Supplementary Figs. S81–S147).	Not reported	Not reported

11. Was the conflict of interest included?	Yes: The authors have no conflicts of interest associated with the material presented in this paper.	Yes: The authors declare that they have no conflict of interests	Yes: No potential conflicts of interest were disclosed.	Yes: The authors declare that they have no competing interests	Not reported

Table c: Systematic review results, search date, number of included studies and notes.

Systematic review identified	Results	Search date; number of studies	Notes
Bae 2016.[65]	Breast cancer risk in Asian women increased with breast density measured using percent density. An overall ES reflecting information from all 6 articles was not calculated, because the number of articles was small and the breast density index varied across articles. For premenopausal women assessed using percent density, the sES was 3.23 (95% CI 2.23, 4.66; two studies). For postmenopausal women assessed using percent density, the sES was 1.62 (95% CI 1.13, 2.32; three studies). The authors concluded that breast cancer risk in Asian women increased with breast density measured using percent density. For pre- and post-menopausal women assessed using Volpara, the summary effect size (sES) was 2.52 (95% CI 1.84, 3.46; one study).	Until December 31, 2015 N=6	Asian women only
Huo 2014.[66]	Mammographic density is associated with increased risk of breast cancer diagnosis. One of the BIRADS studies was reported as showing the OR of an interval cancer for women with dense breasts was 1.62, and the age-adjusted rate ratio was 2.45 for breast cancer incidence (no 95% CI shown). The other BIRADS study was reported as showing that BIRADS IV breasts were more often mammographically occult (no data shown). They found one study using Cumulus and reported that ≥50% density was associated with a 2.63-fold risk of developing breast cancer compared to density <10%; and high density was also associated with ER-positive tumours. The other study of a computer-assisted (semi-automated) method (not stated which) showed that dense area was a better predictor of breast cancer risk than percent density (but no data shown).	Not stated N=37	Very limited information on systematic review methods so scores poorly on AMSTAR
Elias 2014.[67]	Extremely dense breasts on mammography increased the chance of HER2 overexpression (pooled odds ratio [pOR] 1.37; 95% CI, 1.07–1.76).	Through February 2013 N=14	Review focused mainly on HER2 over-expression

Antoni 2013.[68]	The review reported that mammographic density is a strong marker of breast cancer risk. For the eligible study using percent density, the relative risk of ER+ tumours was 1.38 (1.22, 1.57) for low vs. minimal density and the relative risk of ER- tumours was 0.95 (0.67, 1.34). These risks were not shown for the eligible BIRADS studies.	To 5th June 2012 N=19	Q2b by cancer type. Wide age range; no unadjusted analyses; did not report quality assessment of included studies
Cummings 2009[64] and McCormack 2006.[69]	The authors found that breast density was strongly associated with breast cancer: relative risk vs. BIRADS category I was 2.03 (95% CI 1.61, 2.56) for BIRADS II; 2.95 (95% CI 2.32, 3.73) for BIRADS III; and 4.03 (95% CI 3.10, 5.26) for BIRADS IV. For measurement of percent density, vs. <5% dense area, the RR was 1.74 (95% CI 1.50, 2.03) for 5 – 24% density; 2.15 (95% CI 1.87, 2.48) for 25 – 49% density; 2.92 (95% CI 2.55, 3.34) for 50 – 74% density; and 4.20 (95% CI 3.61, 4.89) for >75% density.	January 1, 2004, through January 1, 2008 N=5 additional to those in McCormack 2006	Update of McCormack 2006[69] but does not report the population covered or other details of the included (or excluded) studies

Question 3

Table a: Study design

Yellow highlight = not followed for interval cancers

Study (Country)	Population	Intervention: mammography	Comparator: ultrasound in mammography-negative women	Reference standard	Study design	Limitations
Chang 2015[70] (Korea)	Patients who received mammography (MG) and ultrasound (US) screenings as a prevalence screening examination (n=1526)	Dedicated MG units (Senographic2000 DS units)	Hand-held; high-resolution US units with a 14-15 MHz linear transducer; standardised scanning protocol; bilateral whole breast	Most severe biopsy result within 1 year of screening and clinical follow up at 1 year	Retrospective study	Retrospective, single-institution study performed in a screening center with all examination results interpreted by radiologists specializing in breast imaging. Therefore, the results may not be applicable to other centers with different patient populations or less experience with breast US. Data for cancer detection by US are only available for prevalence screening.

						Although the cancer detection rate and PPV of incidence US screening can be expected to be lower than that of prevalence screening, this is an important consideration because most breast cancer screening examinations involve incidence rather than prevalence screening. MG and US examinations were performed at the same time; the interpretation of mammographic findings can be affected by the US findings. The number of US screen detected cancer was small so it was impossible to find the characteristics of screen detected cancers in this study. Median 47 (range 27-79) years.
Destounis 2015[71] and Destounis 2017[83] (USA)	Screening breast sonography due to notification of dense breast tissue (n=4898 women)	Either a Selenia LoRad or Dimensions unit (Hologic, Inc, Danbury, CT).	Bilateral hand-held US; linear high-frequency transducer; whole breast with standardised protocol using either an iU22 (Philips Healthcare, Bothell, WA) or Acuson S2000 (Siemens Medical Solutions, Malvern, PA) system. All sonograms reviewed by 1 of the radiologists with all prior	Biopsy/surgical excision/histology; no reporting of follow up of test-negative patients	Retrospective electronic chart review	There was a large population of patients with dense tissue pursuing screening sonography who also had additional risk factors. When comparing with our general screening population, we did note that the rate of patients with additional risk factors was quite a bit higher in the population undergoing

			images available for comparison			screening sonography. This factor may have led to a subselection bias. Although we offered screening sonography services to all patients in our screening population identified as having dense breast tissue, those with additional risk factors may have been more inclined to pursue further screening, which could also have had an impact on our study results, as our cancer detection rate could have been higher because of the higher-risk patients. Unrepresentative self-selected sample. Mean 55.8 years 18–35 years: 23 (0.47%) 36–45 years: 855 (17.46%) 46–55 years: 1822 (37.19%) 56–65 years: 1277 (26.07%) 66–75 years: 712 (14.54%) >76 years: 209 (4.27%)
Hwang 2015[72] (Korea)	Asymptomatic women, aged at least 30 years, who underwent mammograms for breast screening (n= 1727)	Bilateral four-view mammograms were obtained using digital mammographic units (Senographe DS, General Electric Medical Systems, Milwaukee, WI, USA; Lorad Selenia, Hologic, Danbury, CT, USA).	Handheld US was performed including bilateral whole breasts and both axillary areas using US units (HDI 5000, Advanced Technology Laboratories, Bothell, WA, USA; IU22, Philips Healthcare, Bothell, WA, USA; Logic 700, General	Pathology and follow-up breast imaging until the year 2011 (around 4 years)	Retrospective cohort study	First, the authors excluded the women who did not visit their institution until December 2011 and the women with mammographic BI-RADS categories 0 and 3. Therefore, there could be more interval cancers which were misclassified as test-

			Electric Medical systems, Milwaukee, WI, USA), equipped with 5–12-MHz linear-array transducers			negatives in the women who underwent mammography plus US screening but were excluded. Second, almost half of the group had baseline screening US and all US examinations were performed by experienced radiologists, which may result in favorable screening US outcomes. The cost of handheld US is not so attractive to patients. Third, the benefit of screening US was only for the detection of early cancers, and did not consider mortality reduction. Multicenter, randomised, prospective studies are required to validate US efficacy as a second line screening tool, and the large-scale data are needed to establish the screening guideline. Participants were self-selected: US was performed in women who requested them, regardless of their risk factors. Median age: 49.5; range 30–76 years. The majority of the women were in their forties (n=763, 44.2%) or in their fifties

						(n=693, 40.1%), and the rest were in their sixties (n=143, 8.3%), 30's (n=107, 6.2%), and seventies (n=21, 1.2%).
Kim 2016[73] (Korea)	Women who underwent screening mammography, who had dense breast defined as BI-RADS density grade 3 (heterogeneously dense) or 4 (extremely dense) at mammography, who had negative findings defined as BI-RADS final assessment category 1 or 2 at mammography, and who had radiologist-performed, hand-held supplemental US examinations performed within 3 months after mammography (n= 3171)	Digital mammography system (Lorad/Hologic Selenia, Lorad/Hologic, Danbury, CT; SENOGRAPHE 2000D, GE Medical Systems, Milwaukee, WI).	Hand-held bilateral whole-breast US was performed with a 12- to 5-MHz linear array transducer (HDI 5000 or iU22, Phillips-Advanced Technology Laboratories, Bothell, WA; Logic 9, GE Medical Systems, Milwaukee, WI). Assessment used the "downgrade criteria": Since March 2010 (the starting year of this study), in order to reduce the false positive rate, the authors have trained their radiologists to classify the following findings as category 2: a complicated cyst 5 mm or smaller which were observed as a circumscribed, homogeneous, and hypoechoic lesion (A) and a circumscribed oval-shaped solid mass 5 mm or smaller without any suspicious US features (B). The 2 criteria for downgrading were selected in consensus after an in-depth discussion	Pathology and 1 year follow up	Retrospective cohort study	This study was retrospectively conducted in a single institution, third-referral center by breast radiologists. Generalisation of the results may be limited for other study populations, and for examinations performed by technologist or less-experienced physicians. Selection bias might have occurred owing to the exclusion of women without follow-up US for at least 1 year. Due to the retrospective nature of the study, the authors could not analyze from the collected data whether the downgrade criteria was properly applied per patient-level by each radiologist. More systematic training programs and quality control programs using videos, still images, or tests are needed to monitor the quality of each radiologist's classification abilities with the downgrade criteria. Further large-scale, multicenter, prospective studies are

			between staff radiologists based on experience and other publications. During the study period, staff radiologists continued to emphasize the downgrade criteria to fellow radiologists at the weekly conference.			needed to validate the effectiveness of the downgrade criteria. Mean age ± standard deviation: 51.2 ± 7.7; range 24–78 years. "Downgrade criteria" not a standard classification.
Klevos 2017[74] (USA)	Asymptomatic women who were reported to have heterogeneously dense or extremely dense breast tissue and negative mammograms (n= 394)	2D digital study on a Selenia - Hologic unit	Hand-held US using a dedicated breast ultrasound unit (GE LOGIC E9) with a high-resolution linear-array transducer (6–15 MHz).	Biopsy result and mammogram at 12 months	Retrospective cohort study	Small population size, which is likely responsible for the fact that no carcinoma was found. Only 32.5% of women underwent the offered supplemental screening bilateral breast ultrasound (may not be representative; ages not stated).
Moon 2015[75] (Korea)	Screening mammography (n=2005 who were BIRADS 1 or 2 on mammography and had screening ultrasound and 1890 BIRADS 1 or 2 on mammography without ultrasound)	Lorad/Hologic Selenia full-field digital mammography and General Electric senograph digital mammography system	US machine: HDI5000 or iU22, Philips-Advanced Technology Laboratories, Bothwell, WA, USA; Logic 9, GE Medical Systems, Milwaukee, WI, USA; and 5-12 or 7-12 MHz linear array transducers. Bilateral whole breasts and axillary areas.	Histopathology from biopsy or surgical excision within 12 months of mammography; clinical follow up for at least 12 months	Retrospective cohort study	Retrospective design; there may be selection bias; only a single round of screening regardless of any previously performed screening was included and the prevalence and incidence of breast cancers were not evaluated separately. Seven radiologists interpreted the screening mammography and performed screening ultrasound; inter-observer variability might impact the results. There was no guideline for recommending and

						performing ultrasound – it was performed according to woman's or clinician's preference, i.e. a self-selected sample undergoing ultrasound. Mean 53.8 (range 40 to 87) years
Tagliafico 2016[76] (Italy)	Asymptomatic women (≥ 38 years old) presenting for mammography screening to public hospital-based radiologic services with dedicated breast imaging were eligible if standard 2D digital mammography was classified as Breast Imaging-Reporting and Data System 22 density categories three (heterogeneously dense) or four (extremely dense) and was negative for BC (n=3231)	Mammography (and tomosynthesis) images were acquired using digital mammography units with tomosynthesis capability (Hologic, Selenia Dimensions; Bedford, MA). Standard 2D-mammography and then 3D-mammography (tomosynthesis) acquisitions were performed in women with dense breasts	Bilateral handheld breast ultrasound was performed using 10 MHz as the lowest maximum frequency of the transducer	Excision histopathology in those who received surgery, or on the basis of the completed assessment inclusive of work-up imaging (with or without core-needle biopsy) in all recalled subjects. No follow up for interval cancers.	Prospective multicenter screening trial of tomosynthesis and ultrasound for adjunct screening in women with dense breasts	These results should be interpreted with caution given that this is an interim analysis, and that the study population comprised women who self-referred to breast screening and who had dense mammograms. Although self-referral to breast screening at the participating centers is intended for women at population (average) risk, we are unable to quantify the risk profile of participating women. However, we can confirm that we did not include women with BRCA gene mutations. Included a modest number of cancers in the interim report. Hence, our incremental CDRs are associated with relatively large CIs; we plan to continue the study to provide more precise estimates at its conclusion.

						Another limitation is that we compared a mix of prevalent and incident ultrasound screening with prevalent tomosynthesis screening, which might give more favorable FP-recall data for ultrasound relative to tomosynthesis. Also, biomarker (eg, estrogen receptor/ progesterone receptor and human epidermal growth factor receptor 2) data were not available for all of the detected cancers. ASTOUND focused on screen-detection measures, and specifically on incremental BC detection; we do not have longer-term data to determine screening benefit because this was not within the scope of the study. The value of adjunct screening could be potentially assessed by follow up of screened subjects and comparing interval cancer rates between those who had adjunct screening and those who did not receive adjunct screening. No follow up for interval cancers.

						Median 51 years (interquartile range, 44 to 78 years; range, 38 to 88 years).
Weigert 2015[77] and Weigert 2017[82] (USA)	Screening ultrasounds performed on women with mammographically normal (BI-RADS 1, normal breasts or BI-RADS 2, stable of known benign finding) but dense breasts (>50% breast density, as determined by the interpreting mammographer) (n= 10282)	Not stated	Ultrasounds using handheld high-resolution transducers (12–5 MHz). None of the sites utilised automated breast ultrasound devices.	Biopsy only; no follow up for interval cancers	Retrospective chart review	The current lack of practice guidelines for screening breast ultrasound results in inconsistency among radiology groups. Ultrasound technologists at some sites document a minimum of a 3, 6, 9, and 12 o'clock image, while at other sites they only record one image if the provider deems the breast is normal. Furthermore, radiologists subjectively determine the degree of breast density when reading screening mammograms and inter-rater reliability is low. Given the study design, the authors do not have enough follow-up data to know how many women developed interval cancers to calculate an accurate NPV or sensitivity. They could not differentiate between women who were receiving screening breast ultrasound for the first time and women who had previously received screening ultrasounds.

						Possible inconsistency of ultrasound performance and interpretation since various independent groups throughout Connecticut were included in the study. In addition, biopsy results could not be obtained for some of the women with ultrasound BI-RADS scores of 4 and 5; it is uncertain if they declined biopsy or went to another location for follow-up. The authors did not include a rigorous follow-up of patients with BIRAD 3 designation to determine if any of those lesions were actually cancers. Of note, only 30% of eligible women returned for the study most likely due to cost and a lack of education. Age not stated
Wilczek 2016[78] (Sweden)	Women invited for breast cancer service screening mammography; age 40 or older; asymptomatic; ACR3 and ACR4 density (n= 1668)	FFDM Microdose Senographe or Senographe DS FFDM	3D ABUS: U-Systems; linear broadband transducer 6-14 MGHZ. All women with suspicious findings on FFDSM or 3D ABUS recalled and had mammography work-up with complementary views and HHUS.	Biopsy or follow up for interval cancers for 2 years	Prospective cohort study	All dedicated breast radiologists involved in the study had to undergo tutorials prior to study initiation, but even so, each one had to familiarize themselves with this new modality, leading to individual learning curves. 3D ABUS was double read only in cases of discussions, while FFDSM was always double

						read. We did not have access to computer-aided detection system for 3D ABUS; such a system could possibly have been of help to reduce reading time and improve early cancer detection. The number of study participants was relatively small in the context of breast screening trials. The study was not designed to detect mortality. Mean (SD) age 49.5 (7.9), range 40-69 years.

Table b: Recall, biopsy and cancer detection rates from the studies found in our update search for ultrasound in mammogram-negative women

Yellow highlight = not followed for interval cancers

				US in mammogram-negative women		
Study (Country)	**USPSTF Quality Rating**	**Breast density**	**Which BIRADS categories (from mammograms) included in study**	**Recall rate per 1000 screens**	**Biopsy rate per 1000 screens**	**Cancer detection rate per 1000 screens**
Chang 2015[70] (Korea)	Fair	Dense or fatty	1 or 2	431/1526 = 282.4/1000	91/1526 = 59.6/1000	5/1526 = 3.3/1000
		Dense only	1 or 2	366/990 = 370/1000		5/990 = 5.1/1000
Destounis 2015[71] and Destounis 2017[83] (USA)	Poor	Dense only	"negative mammograms"	135/5434 = 248/1000	100/4898 = 20.4/1000	18/5434 = 3.3/1000
Hwang 2015[72] (Korea)	Poor	Dense or fatty	1 or 2	100/1727 = 58/1000	25/1727 = 14.5/1000	8/1727 = 4.6/1000
		Dense only	1 or 2	NR	NR	8/1349 = 5.9/1000
Kim 2016[73] (Korea)	Poor	Dense only	1 or 2	831/3171 = 262/1000	147/3171 = 46.4/1000	9/3171 = 2.8/1000

Klevos 2017[74] (USA)	Poor	Dense only	1 or 2	69/394 = 175/1000	26/394 = 66.0/1000	0/394 = 0/1000
Moon 2015[75] (Korea)	Poor	Dense or fatty	1 or 2	623/2005 = 311/1000	90/2005 = 44.9/1000	4/2005 = 2.0/1000
		Dense only	1 or 2	592/1656 = 357/1000	88/1656 = 53.1/1000	3/1656 = 1.8/1000
Tagliafico 2016[76] (Italy)	Poor	Dense only	"negative mammograms"	88/3231 = 27.2/1000	47/3231 = 14.5/1000	23/3231 = 7.1/1,000
Weigert 2015[77] (USA) **Weigert 2017[82] Yr 1** **Year 2** **Year 3** **Year 4**	Poor	Dense only	1 or 2	435/10,282 = 42.3/1000 151/2706 = 55.8/1000 180/3351 = 53.7/1000 148/4128 = 35.9/1000 53/3331 = 15.9/1000	435/10,282 = 42.3/1000 151/2706 = 55.8/1000 180/3351 = 53.7/1000 148/4128 = 35.9/1000 53/3331 = 15.9/1000	24 cancers and 15 high-risk (HR) lesions: total 3.8/1,000; ca 2.3/1,000 11 ca: 4.0/1,000 9 ca/2 HR: tot: 3.3 and ca 2.7/1000 13 ca/2 HR: tot: 3.1 and ca 2.7/1000 10 ca/1 HR: tot: 3.3 and ca 3.0/1000
Wilczek 2016[78] (Sweden)	Poor	Dense only	1 or 2	15/1645 = 9.1/1000	12/1645 = 7.3/1000	4/1645 = 2.4/1000

Table c: Sensitivity, specificity, positive predictive value after recall or after biopsy, and negative predictive value of ultrasound in mammogram-negative women

			US in mammogram-negative women							
Study (Country)	**USPSTF Quality Rating**	**Breast density**	**Recall rate (%)**	**Biopsy recommended (%)**	**Cancer detection rate (per 1000 screens)**	**Sensitivity (%)**	**Specificity (%)**	**PPV1 (%) for recall**	**PPV2 (%) for biopsy**	**NPV (%)**
Chang 2015[70] (Korea)	Fair	Dense or fatty	Recalled (BIRADS 3 or 4 or 5): 431/1526 = 28.24%	Biopsy recommended (BIRADS 4): 104 lesions in 91 women (91/1526 = 5.96%)	3.3 per 1000 screen (95% CI 1.2 to 7.9 per 1000 screens)	5/5 = 100%	1095/1521 = 72.0%	5/431 = 1.2%	5/91 = 5.3%	1095/1095 = 100%

		Dense only	NR	NR	Cancer detection rate 5/990 = 5.1 per 1000 screens (95% CI 1.8 to 12.1 per 1000 screens)	5/5 = 100%	624/985 = 63.4%	5/366 = 1.4%	NR	624/624 = 100%
Destounis 2015[71] and Destounis 2017[83]	Poor	Dense only	135/5434 = 24.8%	100/4898 women = 2.0%	18/5434 ultrasounds = 3.3 per 1000 screens	Not followed for interval cancers	NR	18/135 = 13.3%	18/100 = 18%	Not followed for interval cancers
Hwang 2015[72] (Korea)	Poor	Dense or fatty	100/1727 (5.8%)	25/1727 = 14.5/1000	8/1727 = 4.6 per 1000 cases	8/9 = 88.9%	1626/1718 = 94.6%	8/100 = 8.0%	7/25 = 28.0%	1626/1627 = 99.9%
		Dense only	NR	NR	NR	8/9 = 88.9%	NR	NR	NR	NR
Kim 2016[73] (Korea)	Poor	Dense only	831/3171 = 26.2%	147/3171 = 4.6% (4.1 to 6.8)	9 additional cancers of 3171 screens = 2.8 per 1000 screens, 95% CI 1.3–5.4	9/9 = 100%	2340/3162 = 74.0%	9/831 = 1.1%	9/131 = 6.9%	2340/2340 = 100%
Klevos 2017[74] (USA)	Poor	Dense only	69/394 = 17.5%	26/394 = 6.6%	0	N/A (no cancers found)	N/A	N/A	N/A	N/A
Moon 2015[75] (Korea)	Poor	Dense or fatty	623/2005 = 31.1%	NR	4/2005 = 2.0 per 1000 screens (0.5, 5.1)	4/4 = 100.0%	1382/2001 = 69.1%	4/623 = 0.64%	3/90 = 3.33%	1382/1382 = 100.0%
		Dense only	NR	NR	3/1656 = 1.8 per 1000 screens (0.4, 5.3)	3/3 = 100.0%	1064/1653 = 64.4%	3/592 = 0.51%	2/86 = 2.33%	1064/1064 = 100.0%
Tagliafico 2016[76] (Italy)	Poor	Dense only	88/3231 = 2.72%	47/3231 = 1.45%	23/3231 = 7.1 per 1,000 screens; 95% CI, 4.2 to 10.0	Not followed for interval cancers	98.0%	23/88 = 26.1%	23 per 47 screens (48%; 95% CI, 34.1 to 63.9)	Not followed for interval cancers
Weigert 2015[77]	Poor	Dense only	1310/10,282 = 12.7%	435/10,282 = 4%	2.3/1,000 women screened	Not followed for interval cancers	8,972/9,368 = 96%	Cancers only: 5.5%	Cancers only: 5.5%	Not followed for interval cancers

Weigert 2017[82] Year 1 Year 2 Year 3 Year 4 (USA)								7.3% 5.0% 7.4% 18.9%	7.3% 5.0% 7.4% 18.9%	
Wilczek 2016[78] (Sweden)	Poor	Dense only	0.91%	12/1645 = 0.73%	4/1645 = 2.4/1000	4/9 = 44.4%	1625/1636 = 99.3%	4/15 = 26.7%	4/12 = 33.3%	1625/1630 = 99.7%

Quality assurance guidelines for breast cancer screening radiology from the NHS Breast Screening Programme[1] contain the following radiological quality standards:

Objective	Criteria	Minimum standard	Achievable standard
To minimise the number of women screened who are referred for further tests ‡	The percentage of women who are referred for assessment	(a) Prevalent screen < 10% Incident screen < 7%	(a) Prevalent screen < 7% Incident screen < 5%

‡ 'Further tests' includes all second appointments where procedures (including further views and/or clinical examination) beyond those normally undertaken at first appointment are carried out.

In addition, the expected interval cancer rates after mammography are: 0–24 months: 1.2 invasive cancers per 1000 women screened; 25–36 months: 1.4 per 1000 women screened.

Only three studies[76-78] had a recall rate for ultrasound below 10%.

The rate of benign biopsies (false positives) were as follows:

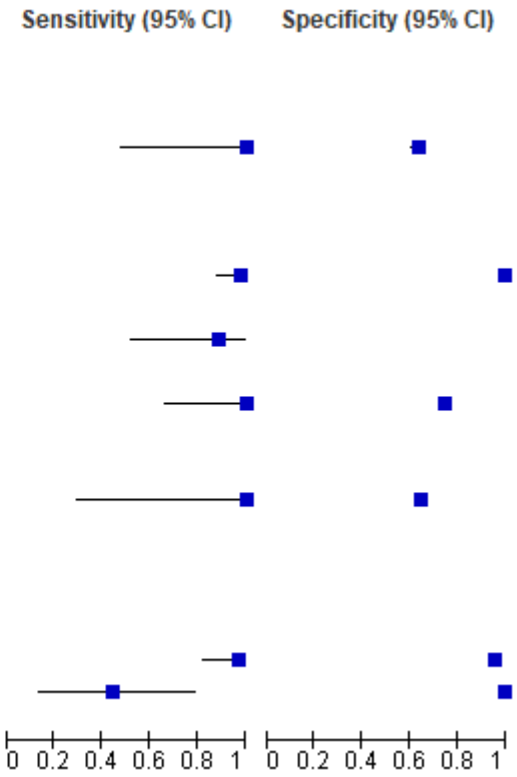
Destounis 2015[71] and Destounis 2017[83] (USA)	17.1/1000

Kim 2016[73] (Korea)	43.6/1000
Klevos 2017[74] (USA)	66.0/1000
Moon 2015[75] (Korea)	51.3/1000
Tagliafico 2016[76] (Italy)	7.4/1000
Weigert 2015[77] (USA)	40.0/1000
Wilczek 2016[78] (Sweden)	4.9/1000

Focusing on the cohort studies reporting data in women with dense breasts only with negative mammography, in which women were followed up for interval cancers, sensitivity ranges from 44% to 100% and specificity from 63% to 99%.

Figure d: Forest plot of sensitivity and specificity of additional ultrasound in mammogram-negative dense breasts

Study	TP	FP	FN	TN	Sensitivity (95% CI)	Specificity (95% CI)
Berg 2012 (high risk)	0	0	0	0	Not estimable	Not estimable
Brancato 2007 (not screening pop)	0	0	0	0	Not estimable	Not estimable
Brem 2015 (no data)	0	0	0	0	Not estimable	Not estimable
Chang 2015	5	361	0	624	1.00 [0.48, 1.00]	0.63 [0.60, 0.66]
Corsetti 2011 (film)	0	0	0	0	Not estimable	Not estimable
Destounis (no data)	0	0	0	0	Not estimable	Not estimable
Girardi 2013 (not data)	0	0	0	0	Not estimable	Not estimable
Giuliano 2013	42	10	1	3365	0.98 [0.88, 1.00]	1.00 [0.99, 1.00]
Hooley 2012 (screen/diag)	0	0	0	0	Not estimable	Not estimable
Hwang 2015	8	0	1	0	0.89 [0.52, 1.00]	Not estimable
Kelly 2010 (high risk)	0	0	0	0	Not estimable	Not estimable
Kim 2016	9	822	0	2340	1.00 [0.66, 1.00]	0.74 [0.72, 0.76]
Klevos 2017 (no data)	0	0	0	0	Not estimable	Not estimable
Leong 2012 (film)	0	0	0	0	Not estimable	Not estimable
Moon 2015	3	589	0	1064	1.00 [0.29, 1.00]	0.64 [0.62, 0.67]
Parris 2013 (not dens)	0	0	0	0	Not estimable	Not estimable
Tagliafico 2016 (no data)	0	0	0	0	Not estimable	Not estimable
Venturini 2013 (not dens)	0	0	0	0	Not estimable	Not estimable
Weigert 15/17 (no data)	0	0	0	0	Not estimable	Not estimable
Weigert 2012	28	401	1	7450	0.97 [0.82, 1.00]	0.95 [0.94, 0.95]
Wilczek 2016	4	11	5	1625	0.44 [0.14, 0.79]	0.99 [0.99, 1.00]
Youk 2011 (film)	0	0	0	0	Not estimable	Not estimable



Question 4

Table a: Characteristics and findings of cost-effectiveness studies investigating supplemental ultrasound in women with mammography-negative dense breasts

Author (Year)	Type of economic evaluation & model	Population studied	Comparators	Methods (perspective, time horizon and discount rate)	Methods (costs, outcomes, ICER and sensitivity analyses)	Results and main conclusions
Giuliano 2013[80]	**EE:** CCA **Model:** None – but simple theoretical calculations	Women with dense breasts in a large screening population in the United States.	**Intervention:** Mammography plus ultrasound **Comparator:** Mammography only	**Study perspective:** Medicare and Medicaid reimbursement **Time horizon:** 1 year **Discount rate:** Not undertaken **Currency/price year:** US$, year not stated	**Outcomes:** additional treatment for missed cancers **Costs:** breast ultrasound, missed cancers, treatments **ICER:** cost per additional treatment for missed cancers **Sensitivity analyses:** Not undertaken	The cost differential for additional treatment between Stage 1 and Stage 2 breast cancer was $10,467. The cost-benefit of early detection of stage 1 disease results in a theoretical per capital annual cost savings of $22.75 per screened patient in the U.S. population, according to their model.
Gray 2017[84] (NB intervention also includes MRI)	**EE**: CUA **Model:** Decision-analytic model (discrete	Women eligible for a national breast screening program (NBSP) in the UK	**Intervention:** Four approaches to stratified NBSP Risk 1 Risk 2	**Perspective:** National health Service **Time horizon:** Lifetime **Discount rate:** 3.5% for both costs and benefits	**Outcomes:** QALYs **Costs:** mammography, follow-up, biopsy, treatments, ultrasound, MRI **ICER:** cost per QALY gained	The risk stratified NBSPs (risk 1 and risk 2) were cost-effective when compared with the current UK NBSP, with ICERs of £16,689 per QALY and £23,924 per QALY, respectively. Stratified NBSP including masking approaches (supplemental screening for

	event simulation)		Masking - current screening approach with supplemental ultrasound offered to women with high breast density. Women with both high breast density and high risk of breast cancer were offered supplemental magnetic resonance imaging (MRI) instead of ultrasound Risk 1 with masking **Comparator:** Current UK NBSP and no screening	**Currency/price year:** UK £ in 2015 prices	**Sensitivity analyses:** One-way and probabilistic sensitivity analyses	women with higher breast density) was not a cost-effective alternative, with ICERs of £212,947 per QALY (masking) and £75,254 per QALY (risk 1 and masking). When compared with no screening, all stratified NBSPs could be considered cost-effective.
Sprague 2015[85]	**EE**: CEA **Model:** 3 micro-simulation models	Women eligible for breast screening in USA. Biennial screening for 50-74 year olds; Annual screening for 40-74 year olds.	**Intervention:** Mammography plus supplemental ultrasound **Comparator:** Mammography alone	**Perspective:** Federal Payer **Time horizon:** Lifetime **Discount rate:** 3% for both costs and benefits **Currency/price year:** US $ in 2013 prices	**Outcomes:** QALYs **Costs:** mammography screening, ultrasound, additional imaging, biopsy, cancer treatment **ICER:** cost per QALY gained **Sensitivity analyses:** One-way sensitivity analyses	Supplemental ultrasound screening for women with dense breasts undergoing screening mammography would substantially increase costs while producing relatively small benefits in breast cancer deaths averted and QALYs gained. The ICER was $325,000 per QALY gained for women with heterogeneously or extremely dense breasts (biennial screening). Restricting supplemental ultrasound screening to women with extremely dense breasts the

						ICER was $246,000 per QALY gained (biennial screening). For annual screening the ICERs were even higher than biennial screening.
Weigert 2012[81]	**EE:** CCA **Model:** None	Women with normal mammograms but dense breasts in the USA	**Intervention:** Mammography plus ultrasound **Comparator:** Mammography alone	**Perspective:** Not stated **Time horizon:** 1 year **Discount rate:** Not undertaken **Currency/price year:** US$, year not stated	**Outcomes:** Number of breast cancers detected **Costs:** average reimbursement by CPT-code and insurance company relating to mammograms, ultrasounds and biopsy's including staff time. **ICER:** Cost per breast cancer found **Sensitivity analyses:** Not undertaken	Using $250 per screening ultrasound and $2,400 per ultrasound-guided biopsy to estimate the costs, the cost per breast cancer found is estimated to be $110,241

EE = economic evaluation; CCA – cost-consequence analysis; CEA = cost-effectiveness analysis; CUA = cost-utility analysis; ICER = incremental cost-effectiveness ratio; QALY = quality-adjusted life year.

Table b: Assessment of the fully-published UK cost-effectiveness study (note intervention includes MRI as well as ultrasound)

Reference	Gray 2017[84]
Interventions and comparators	**Interventions** **Risk 1**: a risk-based stratification defined by the risk algorithm plus density and texture measures. Three strata (with associated screening intervals) were defined by 10-y risks of breast cancer of 1) <3.5% (3-yearly), 2) 3.5%–8% (2-yearly), and 3) >8% (annually) **Risk 2**: a risk-based stratification defined by the same algorithm as risk 1 but with strata defined by dividing the population into thirds on the basis of 10-y risk (tertiles): 1) the lowest risk tertile (3-yearly), 2) the middle tertile (2-yearly), and 3) the highest risk tertile (annually) **Masking** (covering up of tumors in mammograms by dense breast tissue): current screening approach with supplemental ultrasound offered to women with high breast density, defined using Volpara density grade 3 or 4. High risk was defined as >8%

	10-y risk of breast cancer. Women with both high breast density and high risk of breast cancer were offered supplemental magnetic resonance imaging instead of ultrasound. **Risk 1 with masking**: the risk 1 stratification approach together with the strategy described in the masking approach **Comparators** **Current UK NBSP**: women between 50 and 70 y with screening every 3y using mammography **No screening**: no use of mammography in the population for screening purposes; all cancers would present with clinical signs or symptoms
Research question	To identify the incremental costs and consequences of stratified national breast screening programs (stratified NBSPs) and drivers of relative cost-effectiveness.
Study type	Cost-effectiveness analysis
Study population	Women eligible for an NBSP. Mean +/- SD age (y): base case 48.93 +/- 1.09
Institutional setting	National health care service (NHS)
Country/currency	United Kingdom/£. National currency (£) at 2014 prices
Funding source	Part of a European collaborative project called Adapting Breast Cancer Screening Strategy Using Personalised Risk Estimation (ASSURE). The ASSURE project was funded from a collaborative project grant within the FP7-HEALTH-2012- INNOVATION-1 call (project number: 306088).
Analytical perspective	NHS
Effectiveness	Multiple data sources were used: systematic reviews of effectiveness and utility, published studies reporting costs, and cohort studies embedded in existing NBSPs. Mammography and ultrasound sensitivity/specificity etc, interval cancers, survival and effectiveness of MRI referenced. Mammography Sensitivity by tumor size modelled as logistic-type function β1: sets increase with size 1.47 β2: sets sensitivity relative to size 6.51 Maximum sensitivity 0.95% Sensitivity by VDG, used to calculate relative sensitivity given tumor size Sensitivity VDG1 85.0% Sensitivity VDG2 77.6% Sensitivity VDG3 69.0% Sensitivity VDG4 58.6% Recall rate 4.0 per 100 examinations

	False-positive biopsy proportion 2.4% Proportion of screen-detected cancers that are DCIS 20.3% Clinically detected (interval cancers) Cancer size at clinical detection, mean 6.5 doublings (22.62mm) Cancer size at clinical detection, SD 0.535 doublings Survival after breast cancer diagnosis γ NPI 1 -5.413 γ NPI 2 -4.023 γ NPI 3 -2.465 γ Advanced cancer, age <50 y -0.527 γ Advanced cancer, age 50–69 y -0.537 γ Advanced cancer, age ≥70 y -0.849 US cancer detection VDG3/4 incremental cancers detected with supplemental US 3 per 1000 examinations False-positive (recall) rate, US 98 per 1000 examinations Biopsy rate, US 0.4% Assumed same as mammography Proportion cancers detected by supplemental US that are DCIS 21% Assumed same as mammography MRI cancer detection VDG3/4 incremental cancers detected with supplemental US 5 per1000 examinations False-positive (recall) rate, MRI 41.15 per 1000 examinations Biopsy rate, MRI 3.03% Proportion of cancers detected by supplemental MRI that are DCIS 14.3%
Intervention costs	Multiple data sources were used: systematic reviews of effectiveness and utility, published studies reporting costs, and cohort studies embedded in existing NBSPs. Cost data referenced plus expert opinion. Costs Mammography £54 Follow-up, mean £95 Biopsy, mean £160 NPI 1 treatment, mean £11,630 NPI 2 treatment, mean £12,978 NPI 3 treatment, mean £15,405

	Advanced cancer, mean £23,449 Screening ABUS £80 Screening HHUS £80 Screening MRI £220 Stratification process £10.57
Indirect costs	Costs to individual women were excluded from the analysis
Health-state valuations/utilities	Multiple data sources were used: systematic reviews of effectiveness and utility, published studies reporting costs, and cohort studies embedded in existing NBSPs. Utilities referenced Utility Early breast cancer, first year 0.696 Early breast cancer, subsequent years 0.779 Advanced breast cancer, first year 0.685 Advanced breast cancer, subsequent years 0.685
Modelling	A decision-analytic model (discrete event simulation). A de novo model was developed. The conceptualisation process identified that the model required three components to represent: the stratification approach, breast cancer natural history with screening, and the diagnosis and treatment process after a cancer detected by screening. A discrete event simulation (DES) model was used to represent these three components.
Transition probabilities for model	Extensive definitions of various parameters/equations used; also referenced to supplementary material
Time horizon	Lifetime
Discount rates applied in the model for costs and outcomes	3.5% for both costs and benefits (base case) 3.5% for costs and 1.5% for benefits (sensitivity analysis)
Results/analysis: Measure of benefit reported	QALYs
Clinical outcome/benefits estimated for each intervention/strategy	Screening program — QALYs (3.5% discount rate) — Cost (£,2015; 3.5% DR) No screening — 17.6919 — 246 Current UK NBSP — 17.7095 — 654 Risk 1 — 17.7119 — 694 Risk 2 — 17.7181 — 858 Masking — 17.7102 — 809

	Risk 1 and masking 17.7124 870
Synthesis of costs and benefits	Screening program ICER vs. No screening (3.5% DR) UK NBSP (3.5% DR) No screening (1.5% health, 3.5% costs) UK NBSP (1.5% health, 3.5% costs) No screening NA NA NA NA Current UK NBSP £23,197 NA £11,343 NA Risk 1 £22,413 £16,689 £11,363 £11,565 Risk 2 £23,435 £23,924 £11,425 £11,592 Masking £30,772 £212,947 £15,065 £105,412 Risk 1 and masking £30,532 £75,254 £14,707 £33,199 DR = discount rate Masking and risk 1 and masking were dominated by the next alternative (current NBSP and risk 1 stratified NBSP, respectively). The ICERs for the remaining comparisons were £23,197 per QALY for the current NBSP compared with no screening, £16,689 per QALY for risk 1 stratified NBSP compared with masking, and £26,749 for risk 2 stratified NBSP compared with masking and risk 1 stratified NBSP. The risk 1 and risk 2 stratified NBSPs were relatively cost-effective when compared with the current UK NBSP. The masking stratified NBSP does not appear to be a cost-effective alternative when compared with the current UK NBSP. When compared with no screening, all screening programs may be considered cost-effective.
Statistical analysis	Not shown
Sensitivity analysis	One-way sensitivity analyses were used to explore the impact of selected input parameters (referenced to supplementary material). Probabilistic sensitivity analysis (PSA) was performed to quantify the effect of the joint uncertainty.
Scenarios tested in sensitivity analysis	Input parameters and discount rates were varied
Results of the sensitivity analysis	Using an alternative discounting rate of 3.5% for costs and 1.5% for benefits resulted in relatively lower estimated incremental cost-effectiveness ratios (ICERs) for all stratified NBSPs compared with the UK NBSP. One-way sensitivity analysis showed that the reported total costs, total QALYs, and ICERs were sensitive to natural history parameter values (α2 and mean tumour size at clinical detection) and screening performance of mammography (β2). ICERs for stratified programs were moderately sensitive to the cost of stratification although costs would need to be several times the base-case value for ICERs to increase beyond a threshold of £30,000 per QALY. In all alternative programs, total costs were sensitive to the treatment cost parameters; varying these parameters, however, did not greatly change the ICERs compared with the base case. Estimates of total QALYs were sensitive to the utility weights for cancer states; varying utility weights

	moderately altered the ICERs of stratified programs compared with the NBSP. The results were relatively insensitive (within the ranges tested) to the probability of recall, costs of MRI, the relative sensitivity of mammography by VDG group, and US/MRI additional cancer detection rate.
Conclusions/implications	A risk stratified NBSP is potentially a cost-effective use of health care resources when compared with the current UK NBSP.
Implications of the evaluation for practice	This early model-based cost-effectiveness analysis provides indicative evidence for decision makers to understand the key drivers of costs and QALYs for exemplar stratified NBSP. Key drivers of cost-effectiveness were discount rate, natural history model parameters, mammographic sensitivity, and biopsy rates for recalled cases. A key assumption was that the risk model used in the stratification process was perfectly calibrated to the population.

Table c: Quality assessment of studies using CHEERS

CHEERS checklist[33]	**Giuliano 2013[80]**	**Gray 2017[84]**	**Sprague 2015[85]**	**Weigert 2012[81]**
Title and abstract				
1 Title: Identify the study as an economic evaluation, or use more specific terms such as ``cost-effectiveness analysis``, and describe the interventions compared.	N	Y	Y	N
2 Abstract: Provide a structured summary of objectives, methods including study design and inputs, results including base case and uncertainty analyses, and conclusions.	*	Y	Y	N
Introduction				
3 Background & objectives: Provide an explicit statement of the broader context for the study. Present the study question and its relevance for health policy or practice decisions.	Y	Y	Y	*
Methods				
4 Target Population and Subgroups: Describe characteristics of the base case population and subgroups analysed including why they were chosen.	Y	Y	Y	N

5 Setting and Location: State relevant aspects of the system(s) in which the decision(s) need(s) to be made.	N	Y	Y	Y
6 Study perspective: Describe the perspective of the study and relate this to the costs being evaluated.	*	Y	Y	N
7 Comparators: Describe the interventions or strategies being compared and state why they were chosen.	Y	Y	Y	Y
8 Time Horizon: State the time horizon(s) over which costs and consequences are being evaluated and say why appropriate.	*	Y	Y	Y
9 Discount Rate: Report the choice of discount rate(s) used for costs and outcomes and say why appropriate.	N	Y	Y	N
10 Choice of Health Outcomes: Describe what outcomes were used as the measure(s) of benefit in the evaluation and their relevance for the type of analysis performed.	*	Y	Y	*
11a Measurement of Effectiveness - Single Study-Based Estimates: Describe fully the design features of the single effectiveness study and why the single study was a sufficient source of clinical effectiveness data.	*	N/A	N/A	*
11b Measurement of Effectiveness - Synthesis-based Estimates: Describe fully the methods used for identification of included studies and clinical effectiveness data synthesis of clinical effectiveness data.	N/A	Y	Y	N/A
12 Measurement and Valuation of Preference-based Outcomes: If applicable, describe the population and methods used to elicit preferences for health outcomes.	N	*	*	N/A
13a Estimating Resources and Costs - Single Study-based Economic evaluation: Describe approaches used to estimate resource use associated with the alternative interventions. Describe primary or secondary research methods for valuing	N	N/A	N/A	*

each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs.				
13b Estimating Resources and Costs - Model-based Economic Evaluation: Describe approaches and data sources used to estimate resource use associated with model health states. Describe primary or secondary research methods for valuing each resource item in terms of its unit cost. Describe any adjustments made to approximate to opportunity costs.	N/A	Y	*	N/A
14 Currency, Price Date and Conversion: Report the dates of the estimated resource quantities and unit costs. Describe methods for adjusting estimated unit costs to the year of reported costs if necessary. Describe methods for converting costs into a common currency base and the exchange rate.	N	Y	Y	N
15 Choice of Model: Describe and give reasons for the specific type of decision-analytic model used. Providing a figure to show model structure is strongly recommended.	N	Y	*	N
16 Assumptions: Describe all structural or other assumptions underpinning the decision-analytic model.	N	Y	Y	N
17 Analytic Methods: Describe all analytic methods supporting the evaluation. This could include methods for dealing with skewed, missing or censored data, extrapolation methods, methods for pooling data, approaches to validate a model, & methods for handling population heterogeneity and uncertainty.	N	Y	Y	N
Results				
18 Study parameters: Report the values, ranges, references, and if used, probability distributions for all parameters. Report reasons or sources for distributions used to represent	N	Y	Y	N

uncertainty where appropriate. We strongly recommend the use of a table to show the input values.				
19. Incremental costs and outcomes: For each intervention, report mean values for the main categories of estimated costs and outcomes of interest, as well as mean differences between the comparator groups. If applicable, report incremental cost-effectiveness ratios.	*	Y	Y	*
20a Characterizing Uncertainty - Single study-based economic evaluation: Describe the effects of sampling uncertainty for the estimated incremental cost and incremental effectiveness, parameters together with the impact of methodological assumptions.	N	N/A	N/A	N
20b Characterizing Uncertainty - Model-based economic evaluation: Describe the effects on the results of uncertainty for all input parameters, and uncertainty related to the structure of the model and assumptions.	N/A	Y	*	N/A
21 Characterizing Heterogeneity: If applicable, report differences in costs, outcomes or in cost-effectiveness that can be explained by variations between subgroups of patients with different baseline characteristics or other observed variability in effects that are not reducible by more information.	N	N	N	N/A
Discussion				
22 Study Findings, Limitations, Generalizability, and Current Knowledge: Summarize key study findings and describe how they support the conclusions reached. Discuss limitations and the generalizability of the findings and how the findings fit with current knowledge.	Y	Y	Y	Y
Other				

23 Source of Funding: Describe how the study was funded and the role of the funder in the identification, design, conduct and reporting of the analysis. Describe other non-monetary sources of support.	N	Y	Y	N
24 Conflicts of Interest: Describe any potential for conflict of interest among study contributors in accordance with journal policy. In the absence of a journal policy, we recommend authors comply with International Committee of Medical Journal Editors' recommendations	N	N	Y	Y

Key: y = yes, n = no, N/A = not applicable and * = partially completed

Appendix 7 Criteria for appraising the viability, effectiveness and appropriateness of a screening programme

UK National Screening Committee criteria for screening programmes published in 2015[28] are:

1. The condition

1. The condition should be an important health problem as judged by its frequency and/or severity. The epidemiology, incidence, prevalence and natural history of the condition should be understood, including development from latent to declared disease and/or there should be robust evidence about the association between the risk or disease marker and serious or treatable disease.

2. All the cost-effective primary prevention interventions should have been implemented as far as practicable.

3. If the carriers of a mutation are identified as a result of screening the natural history of people with this status should be understood, including the psychological implications.

2. The test

4. There should be a simple, safe, precise and validated screening test.

5. The distribution of test values in the target population should be known and a suitable cut-off level defined and agreed.

6. The test, from sample collection to delivery of results, should be acceptable to the target population.

7. There should be an agreed policy on the further diagnostic investigation of individuals with a positive test result and on the choices available to those individuals.

8. If the test is for a particular mutation or set of genetic variants the method for their selection and the means through which these will be kept under review in the programme should be clearly set out.

3. The intervention

9. There should be an effective intervention for patients identified through screening, with evidence that intervention at a pre-symptomatic phase leads to better outcomes for the screened individual compared with usual care. Evidence relating to wider benefits of screening, for example those relating to family members, should be taken into account where available. However, where there is no prospect of benefit for the individual screened then the screening programme should not be further considered.

10. There should be agreed evidence based policies covering which individuals should be offered interventions and the appropriate intervention to be offered.

4. The screening programme

11. There should be evidence from high quality randomised controlled trials that the screening programme is effective in reducing mortality or morbidity. Where screening is aimed solely at providing information to allow the person being screened to make an "informed choice" (such as Down's syndrome or cystic fibrosis carrier screening), there must be evidence from high quality trials that the test accurately measures risk. The information that is provided about the test and its outcome must be of value and readily understood by the individual being screened.

12. There should be evidence that the complete screening programme (test, diagnostic procedures, treatment/ intervention) is clinically, socially and ethically acceptable to health professionals and the public.

13. The benefit gained by individuals from the screening programme should outweigh any harms, for example from overdiagnosis, overtreatment, false positives, false reassurance, uncertain findings and complications.

14. The opportunity cost of the screening programme (including testing, diagnosis and treatment, administration, training and quality assurance) should be economically balanced in relation to expenditure on medical care as a whole (value for money). Assessment against this criteria should have regard to evidence from cost benefit and/or cost effectiveness analyses and have regard to the effective use of available resource.

5. Implementation criteria

15. Clinical management of the condition and patient outcomes should be optimised in all health care providers prior to participation in a screening programme.

16. All other options for managing the condition should have been considered (such as improving treatment or providing other services), to ensure that no more cost effective intervention could be introduced or current interventions increased within the resources available.

17. There should be a plan for managing and monitoring the screening programme and an agreed set of quality assurance standards.

18. Adequate staffing and facilities for testing, diagnosis, treatment and programme management should be available prior to the commencement of the screening programme.

19. Evidence-based information, explaining the purpose and potential consequences of screening, investigation and preventative intervention or treatment, should be made available to potential participants to assist them in making an informed choice.

20. Public pressure for widening the eligibility criteria for reducing the screening interval, and for increasing the sensitivity of the testing process, should be anticipated. Decisions about these parameters should be scientifically justifiable to the public.

6. References

- Department of Health, Screening of pregnant women for hepatitis B and immunisation of babies at risk. London: Dept of Health, 1998 (Health Service Circular : HSC 1998/127).

- Wilson JMG, Jungner G. Principles and practice of screening for disease. Public Health Paper Number 34. Geneva: WHO, 1968.
- Cochrane AL. Holland WW. Validation of screening procedures. Br Med Bull. 1971, 27, 3.
- Sackett DL, Holland WW. Controversy in the detection of disease. Lancet 1975;2:357-9.
- Wald NJ (Editor). Antenatal and Neonatal screening. Oxford University Press, 1984.
- Holland WW, Stewart S. Screening in Healthcare. The Nuffield Provincial Hospitals Trust, 1990.
- Gray JAM. Dimensions and definitions of screening. Milton Keynes: NHS Executive Anglia and Oxford, Research and Development.
- Angela Raffle/Muir Gray Screening Evidence and Practice, Oxford University Press 2007.